

# Test–retest reliability of self-reported diabetes diagnosis in the Norwegian Women and Cancer Study: A population-based longitudinal study (n =33,919)

SAGE Open Medicine  
4: 2050312115622857  
© The Author(s) 2016  
Reprints and permissions:  
sagepub.co.uk/journalsPermissions.nav  
DOI: 10.1177/2050312115622857  
smo.sagepub.com  


Mashhood Ahmed Sheikh, Eiliv Lund and Tonje Braaten

## Abstract

**Objective:** Self-reported information from questionnaires is frequently used in epidemiological studies, but few of these studies provide information on the reproducibility of individual items contained in the questionnaire. We studied the test–retest reliability of self-reported diabetes among 33,919 participants in Norwegian Women and Cancer Study.

**Methods:** The test–retest reliability of self-reported type 1 and type 2 diabetes diagnoses was evaluated between three self-administered questionnaires (completed in 1991, 1998, and 2005 by Norwegian Women and Cancer participants) by kappa agreement. The time interval between the test–retest studies was ~7 and ~14 years. Sensitivity of the kappa agreement for type 1 and type 2 diabetes diagnoses was assessed. Subgroup analysis was performed to assess whether test–retest reliability varies with age, body mass index, physical activity, education, and smoking status.

**Results:** The kappa agreement for both types of self-reported diabetes diagnoses combined was good ( $\geq 0.65$ ) for all three test–retest studies (1991–1998, 1991–2005, and 1998–2005). The kappa agreement for type 1 diabetes was good ( $\geq 0.73$ ) in the 1991–2005 and the 1998–2005 test–retest studies, and very good (0.83) in the 1991–1998 test–retest study. The kappa agreement for type 2 diabetes was moderate (0.57) in the 1991–2005 test–retest study and good ( $\geq 0.66$ ) in the 1991–1998 and 1998–2005 test–retest studies. The overall kappa agreement in the 1991–1998 test–retest study was stronger than in the 1991–2005 test–retest study and the 1998–2005 test–retest study. There was no clear pattern of inconsistency in the kappa agreements within different strata of age, BMI, physical activity, and smoking. The kappa agreement was strongest among the respondents with 17 or more years of education, while generally it was weaker among the least educated group.

**Conclusion:** The test–retest reliability of the diabetes was acceptable and there was no clear pattern of inconsistency in the kappa agreement stratified by age, body mass index, physical activity, and smoking. The study suggests that self-reported diabetes diagnosis from middle-aged women enrolled in the Norwegian Women and Cancer Study is reliable.

## Keywords

Type 2 diabetes, type 1 diabetes, metabolic syndrome, kappa, test–retest reliability, reproducibility, questionnaires, Norway, Norwegian Women and Cancer, Kvinner og kreft

Date received: 1 September 2015; accepted: 17 November 2015

## Introduction

Epidemiological studies often rely on self-reported information, as this renders the costs of data collection lower than that of clinical studies.<sup>1</sup> However, the validity and reliability of the instruments used for data collection are often not reported.<sup>2</sup>

Commonly, the Cohen's kappa coefficient is used to determine inter-rater agreement for disease (or other categorical outcomes) by comparing self-reported information against a gold standard (diagnostic test, medical

records, physiological measures, etc.). Previous validation studies of self-reported diabetes diagnosis have

Department of Community Medicine, University of Tromsø, Tromsø, Norway

### Corresponding author:

Mashhood Ahmed Sheikh, Department of Community Medicine, University of Tromsø, N-9037 Tromsø, Norway.  
Email: mashhood.a.sheikh@uit.no



Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (<http://www.creativecommons.org/licenses/by/3.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

indicated that diabetes is reported more accurately than other illnesses or diseases.<sup>3–10</sup>

The Cohen's kappa coefficient can also be used to analyze the test–retest reliability of an instrument. Many studies from Norway have used self-reported information from questionnaires as the principle tool, but few<sup>11–43</sup> of them have provided information on the reproducibility of the individual items and instruments therein. It is important to establish that respondents with different socio-demographic background, and age groups have understood the questions in a similar manner. Test–retest reliability is assessed by measuring the responses of the same study sample to an identical question at two or more points in time.<sup>44</sup> These responses are then compared to establish the reliability of the instrument. The chi-square ( $\chi^2$ ) test for independence is not appropriate for assessing test–retest reliability since it does not take into account that the data are paired (i.e. different measurements for the same individual).

Previous studies using self-reported data from interviews have studied the test–retest reliability of self-reported diabetes diagnosis, with inconsistent kappa agreements.<sup>45–50</sup> Since type 2 diabetes typically affects people aged 40 years and over,<sup>51,52</sup> it is possible to differentiate between the test–retest reliability of self-reported type 1 and type 2 diabetes diagnoses using information on age at diagnosis. No previous study was found that assessed the test–retest reliability for either type 1 or type 2 diabetes separately.

The Norwegian Women and Cancer (NOWAC) Study<sup>53</sup> is a prospective cohort study in which women reported diabetes diagnosis and age at diagnosis in three separate questionnaires. If a woman accurately reported her diabetes diagnosis in one study, she is expected to report the same in a subsequent study. This assumption underlies our test–retest reliability analysis. The aim of this study was to assess the test–retest reliability of self-reported diabetes diagnosis, as well as that of type 1 and type 2 diabetes diagnoses separately. Furthermore, the large sample size permits subgroup analyses and sensitivity analysis. We examined whether test–retest reliability varies with age, body mass index (BMI), physical activity, education, and smoking status.

## Methods

### *Study cohort and sampling*

The NOWAC Study is a prospective nationwide study which started in 1991,<sup>54,55</sup> and contains data from 170,000 women. Participants were randomly selected from the National Population Register of Norway. The external validity of the study<sup>56</sup> and validity of some measures<sup>57–59</sup> have been published elsewhere. NOWAC Study participants are assumed to be representative of the female Norwegian population in the corresponding age groups.<sup>56</sup> The detailed characteristics of the participants are described

elsewhere,<sup>56</sup> and the updated information on the NOWAC Study is accessible on its website.<sup>54</sup>

Of the 170,000 women enrolled in the NOWAC Study, 33,919 women completed all of three questionnaires sent in 1991, 1998, and 2005. The general characteristics of the study sample and the association between BMI and type 2 diabetes in this sample are described elsewhere.<sup>52</sup>

### *Questionnaire and classification*

**Diabetes.** Information on diabetes diagnosis was collected by means of the same question in all three questionnaires (1991, 1998, and 2005): “Have you had any of the following diseases?” The list of options included diabetes. Age at diagnosis was measured with the subsequent question, “If yes, at what age was it first discovered?” For the purposes of this study, only participants who reported having diabetes and provided their age at diagnosis were defined as diabetes cases. If participants reported they gave birth to a child either the same year they were diagnosed with diabetes, or in the year preceding child birth, it was assumed that they had gestational diabetes, and they were excluded from the analysis. Final numbers of diabetes cases included in analyses are given in Tables 2–4. Participants with missing values on diabetes diagnosis and age at diagnosis were excluded.

Using the responses to the questions on diabetes and age at diagnosis, different variables for diabetes diagnosis, and separate variables for type 1 and type 2 diabetes, were created. Since type 2 diabetes typically affects people aged 40 years or over,<sup>51,52</sup> we classified only those aged 40 years or over as having type 2 diabetes. Women who were diagnosed with diabetes at or before age 39 years were categorized as having type 1 diabetes (excluding those with gestational diabetes). Participants with type 1 and type 2 diabetes were classified separately by the above-mentioned criteria for the 1991 test study, the 1998 test study, the 1998 retest study for comparison against 1991 test study, the 2005 retest study for comparison against the 1991 test study, and the 2005 retest study for comparison with the 1998 test study.

Diabetes cases in the 1991 and 1998 test studies were defined as those who reported having diabetes, and their age at diagnosis in the corresponding questionnaires. One respondent to the 1998 questionnaire fulfilled the criteria for both gestational diabetes and type 2 diabetes and was finally classified as having gestational diabetes only.

**Diabetes in the 1998 retest study (for comparison against the 1991 test study).** Diabetes cases in the 1998 retest study, for comparison against the 1991 test study were defined as those with diabetes from the 1998 test study, provided they reported a date of diagnosis prior to 1992. The same criteria were applied to women with type 1 or type 2 diabetes. One woman in the 1998 retest study fulfilled the criteria both for gestational and type 2 diabetes and was finally classified as having gestational diabetes only.

*Diabetes in the 2005 retest study (for comparison against the 1991 test study).* Diabetes cases from the 2005 retest study, for comparison against 1991 test study, were defined as participants who reported a diabetes diagnosis in the 2005 questionnaire, provided they reported a date of diagnosis prior to 1992. The same criteria were applied to women with type 1 or type 2 diabetes.

*Diabetes in the 2005 retest study (for comparison against 1998 test study).* Diabetes cases from the 2005 retest study, for comparison against the 1998 test study, were defined as participants with self-reported diabetes in the 2005 questionnaire, provided that they reported a date of diagnosis prior to 1999. The same criteria were applied to women with type 1 or type 2 diabetes.

*Covariates.* Self-reported information on height and weight from 1998 study was used to calculate BMI (kg/m<sup>2</sup>). BMI was categorized into three groups: normal weight (BMI: <25 kg/m<sup>2</sup>), overweight (BMI: 25–29.9 kg/m<sup>2</sup>), and obese (BMI: ≥30 kg/m<sup>2</sup>). Smoking status was derived from the replies to two questions in the 1998 questionnaire: “Have you ever smoked?” (yes, no) and “Do you smoke on a daily basis at the moment?” (yes, no). Women who answered “no” to the former were categorized as “never smokers.” Those who answered “yes” to the former, and “no” to the latter, were categorized as “former smokers,” and those who answered “yes” to both questions were categorized as “current smokers.” A 10-category scale measured the level of self-reported physical activity in the 1998 questionnaire, the validity of which has been reported previously.<sup>21</sup> Responses to questions about physical activity were used to assign a category of physical activity: low [1–3], medium [4–7], and high [8–10]. Education (duration in years) was categorized into four groups: primary/intermediate (0–9), secondary (10–12), university (13–16), and postgraduate and above (17+). Age (years) was categorized in four groups with 5-year interval.

### Statistical analysis

Statistical analysis was performed with SAS version 9.2 and Stata version 13.1. Means (standard deviation (SD)) were estimated for all continuous variables, and the percentage of participants in each category was calculated for all categorical variables. General characteristics of the data are presented as frequencies, percentages, and means with SDs, respectively (Table 1). Variables for all diabetes diagnoses, as well as for type 1 and type 2 diabetes separately, were constructed, and the kappa agreement for the two types of diabetes was calculated for the 1991–1998 test–retest study, the 1991–2005 test–retest study, and 1998–2005 test–retest study, respectively. The kappa coefficients summarize the total agreement beyond that expected by chance. 95% confidence intervals (CIs) for kappa statistic were estimated with analytical

method<sup>60</sup> in Stata.<sup>61</sup> Established benchmarks<sup>62,63</sup> for rating the strength of kappa agreements as poor (<0.20), fair (>0.20 to ≤0.40), moderate (>0.40 to ≤0.60), good (>0.60 to ≤0.80), and very good (>0.80 to ≤1.00) were used.

Consistency (%) was calculated as

$$\text{Consistency (\%)} = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100$$

### Sensitivity analysis

Since self-reported age at diagnosis was used as the only discriminative criterion for distinguishing between type 1 and type 2 diabetes, sensitivity analysis was performed by restricting age at diagnosis <35 years for type 1 diabetes and age at diagnosis >44 years for type 2 diabetes (Table 5). Those reporting age at diagnosis 35–44 were excluded for the purpose of assessing sensitivity of the kappa agreements (Table 5).

### Subgroup analysis

Subgroup analysis was performed to assess the consistency of the kappa agreement across stratas of the covariates (Tables 6–10).

### Ethical approval

The NOWAC Study was approved by the Regional Committee for Medical and Health Research Ethics. All participating women gave written informed consent.

## Results

Table 1 presents the general characteristics of the study sample. Among the 33,919 women participating in 1991, 1998, and 2005 study, the age distribution was between 40 and 59 (mean: 47.7±4.3) in 1998. Majority (64.6%) of the respondents had normal weight (BMI: <25 kg/m<sup>2</sup>). Almost 40.3% of the respondents had some university education or more. Most (75.5%) of the respondents were classified as having medium level of physical activity. In this study sample, 28.2% were classified as being current smoker, while 31.2% were classified as being former smokers.

Table 2 presents the self-reported diabetes diagnosis in 1991 study, and 1998 study by self-reported age at diagnosis in respective studies. Majority (56%) of the self-reported diabetics reported age at diagnosis as 30 years or over in 1991 study, while over 64.7% reported age at diagnosis as 40 years or over in the 1998 study. This may partly be due to the aging cohort itself.

Tables 3 and 4 present the kappa statistics for the test–retest studies. The agreement for all self-reported diabetes

**Table 1.** General characteristics of the study sample (n = 33,919).

	Cohort n = 33,919	
	N (%)	Mean (SD)
Age (years)		47.7 (4.3)
40–44	9926 (29.3)	
45–49	11,382 (33.6)	
50–54	10,849 (32.0)	
55–59	1762 (5.2)	
BMI <sup>a</sup>		24.4 (3.8)
Normal weight (<25 kg/m <sup>2</sup> )	21,553 (64.6)	
Overweight (25–29.9 kg/m <sup>2</sup> )	9106 (27.3)	
Obese (≥30 kg/m <sup>2</sup> )	2709 (8.1)	
Education level (duration in years) <sup>a</sup>		12.5 (3.2)
Primary/intermediate (0–9)	6736 (20.1)	
Secondary (10–12)	12,102 (36.1)	
University (13–16)	10,226 (30.5)	
Postgraduate and above (17+)	4460 (13.3)	
Physical activity level <sup>a</sup>		5.6 (1.7)
Low	3686 (11.5)	
Medium	24,229 (75.5)	
High	4186 (13.0)	
Smoking status		
Never smoker	13,763 (40.6)	
Former smoker	10,582 (31.2)	
Current smoker	9574 (28.2)	

SD: standard deviation; BMI: body mass index.

<sup>a</sup>Cohort size was 33,919, but because of missing values, the numbers for some variables do not add up to 33,919.

**Table 2.** Self-reported diabetes diagnosis in 1991 and 1998 test studies by self-reported age at diagnosis.

	Age groups	Diabetic in 1991 <sup>a</sup>	Diabetic in 1998 <sup>b</sup>
		n (%)	n (%)
Age at diagnosis	0–4	5 (3.4) <sup>c</sup>	4 (1.3) <sup>d</sup>
	5–9	10 (6.8) <sup>c</sup>	6 (1.9) <sup>d</sup>
	10–14	18 (12.2) <sup>c</sup>	17 (5.4) <sup>d</sup>
	15–19	10 (6.8) <sup>c</sup>	11 (3.5) <sup>d</sup>
	20–24	7 (4.7) <sup>c</sup>	12 (3.8) <sup>d</sup>
	25–29	15 (10.1) <sup>c</sup>	12 (3.8) <sup>d</sup>
	30–34	23 (15.5) <sup>c</sup>	19 (6.0) <sup>d</sup>
	35–39	23 (15.5) <sup>c</sup>	30 (9.5) <sup>d</sup>
	40–44	25 (16.9) <sup>c</sup>	75 (23.8) <sup>d</sup>
	45–49	12 (8.1) <sup>c</sup>	70 (22.2) <sup>d</sup>
	50–54	–	59 (18.7) <sup>d</sup>
Total	148 (100.0)	315 (100.0)	

<sup>a</sup>Diabetes cases in the 1991 test study were defined as those who reported having diabetes, and their age at diagnosis in the 1991 study.

<sup>b</sup>Diabetes cases in the 1998 test study were defined as those who reported having diabetes, and their age at diagnosis in the 1998 study. One respondent to the 1998 questionnaire fulfilled the criteria for both gestational diabetes and type 2 diabetes and was excluded.

<sup>c</sup>N and % of respondents reporting age at diagnosis in 1991 study.

<sup>d</sup>N and % of respondents reporting age at diagnosis in 1998 study.

diagnoses in the 1991–1998 test–retest study was 0.75 (95% CI: 0.70–0.81), while it was 0.70 (95% CI: 0.66–0.74) in the 1998–2005 test–retest study. The kappa agreement for all self-reported diabetes diagnoses in the 1991–2005 test–retest study was 0.65 (95% CI: 0.58–0.71) (Table 3).

Table 4 shows the kappa agreement for the three test–retest studies separately for the two types of diabetes. The kappa agreement for type 1 diabetes was very good in the 1991–1998 test–retest study (kappa=0.83, 95% CI: 0.76–0.89), while it was good in the 1991–2005 test–retest study (kappa=0.76, 95% CI: 0.68–0.84), and the 1998–2005 test–retest study (kappa=0.73, 95% CI: 0.66–0.81). The kappa agreement for type 2 diabetes was good in the 1991–1998 test–retest study (kappa=0.67, 95% CI: 0.55–0.79), and in the 1998–2005 test–retest study (kappa=0.66, 95% CI: 0.59–0.72), while it was moderate in the 1991–2005 test–retest study (kappa=0.57, 95% CI: 0.43–0.71) (Table 4). The overall kappa agreement in the 1991–1998 test–retest study was stronger than in the 1991–2005 test–retest study and the 1998–2005 test–retest study (Table 4).

Table 5 presents the sensitivity of the kappa agreements by classifying those reporting age at diagnosis less than 35, as diagnosed with type 1 diabetes. While, classifying those reporting age at diagnosis greater than 44 as diagnosed with type 2 diabetes. The kappa agreements remained moderate to good for type 1 diabetes, while the kappa agreements for type 2 diabetes were fair to good (Table 5).

Tables 6–10 present the kappa agreement for diabetes stratified by age, BMI, physical activity, education, and smoking status. There was no clear pattern of inconsistency in the kappa agreements within different strata of age, BMI, physical activity, and smoking (Table 6–8 and 10). However, the stratified analysis by the level of education shows that the kappa agreement is strongest among the most educated group (Table 9) in all the test–retest comparisons, while generally it was weaker among the least educated group.

## Discussion

In this study, we analyzed the test–retest reliability of self-reported diabetes diagnosis in a large sample of middle-aged women in Norway. We observed that the agreement was good for all diabetes diagnoses combined in all three test–retest studies. The weakest agreement was found in the 1991–2005 test–retest study. This was to be expected, as the time interval between these studies was the longest. These results also suggest that other confounding factors may have affected self-reported diabetes diagnosis in the 1991–1998, or 1998–2005 test–retest studies, as the agreement in these periods was expected to be more similar. The fact that diabetes diagnosis may change over time could have contributed to the decreasing agreement observed between the 1991–1998 test–retest study and the 1991–2005 test–retest study. However, looking at the two types of diabetes separately revealed some differences in the kappa agreement. The

**Table 3.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes).

	Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
Diabetes	Cases in 1991 (n) 148	Cases in 1998 (n) 151	113/148 (76.4)	0.75 (0.70–0.81)
	Cases in 1991 (n) 148	Cases in 2005 (n) 130	90/148 (60.8)	0.65 (0.58–0.71)
	Cases in 1998 (n) 315	Cases in 2005 (n) 282	209/315 (66.3)	0.70 (0.66–0.74)

CI: confidence interval.

$$^a \text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100$$

**Table 4.** Kappa agreements for self-reported type 1 and type 2 diabetes diagnoses.

Diabetes type	Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
Type 1 diabetes <sup>b</sup>	Cases in 1991 (n) 111	Cases in 1998 (n) 103	83/111 (74.7)	0.83 (0.76–0.89)
	Type 2 diabetes <sup>c</sup> 37	48	29/37 (78.4)	0.67 (0.55–0.79)
Type 1 diabetes <sup>b</sup>	Cases in 1991 (n) 111	Cases in 2005 (n) 88	64/111 (57.6)	0.76 (0.68–0.84)
	Type 2 diabetes <sup>c</sup> 37	42	21/37 (56.6)	0.57 (0.43–0.71)
Type 1 diabetes <sup>b</sup>	Cases in 1998 (n) 111	Cases in 2005 (n) 97	70/111 (63.1)	0.73 (0.66–0.81)
	Type 2 diabetes <sup>c</sup> 204	185	125/204 (61.3)	0.66 (0.59–0.72)

CI: confidence interval.

$$^a \text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100$$

<sup>b</sup>Type 1 diabetes were classified as those reporting age at diagnosis <40 years.

<sup>c</sup>Type 2 diabetes were classified as those reporting age at diagnosis >39 years.

**Table 5.** Sensitivity analysis of kappa agreements for self-reported type 1 (age at diagnosis: <35 years) and type 2 diabetes diagnoses (age at diagnosis: >44 years).

Diabetes type	Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
Type 1 diabetes <sup>b</sup>	Cases in 1991 (n) 88	Cases in 1998 (n) 81	68/88 (77.3)	0.80 (0.65–0.95)
	Type 2 diabetes <sup>c</sup> 12	15	6/12 (50.0)	0.52 (0.27–0.77)
Type 1 diabetes <sup>b</sup>	Cases in 1991 (n) 88	Cases in 2005 (n) 74	54/88 (61.4)	0.69 (0.51–0.88)
	Type 2 diabetes <sup>c</sup> 12	12	3/12 (25.0)	0.33 (0.05–0.61)
Type 1 diabetes <sup>b</sup>	Cases in 1998 (n) 81	Cases in 2005 (n) 74	57/81 (70.4)	0.60 (0.38–0.81)
	Type 2 diabetes <sup>c</sup> 129	123	75/129 (58.1)	0.63 (0.56–0.70)

CI: confidence interval.

$$^a \text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100$$

<sup>b</sup>Only those reporting age at diagnosis <35 years were included.

<sup>c</sup>Only those reporting age at diagnosis >44 years were included.

**Table 6.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes) stratified by age groups.

		Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
		Cases in 1991 (n)	Cases in 1998 (n)		
Age	40–44	38	31	27/38 (71.1)	0.78 (0.67–0.89)
	45–49	39	38	27/39 (69.2)	0.70 (0.58–0.82)
	50–54	59	65	48/59 (81.4)	0.77 (0.69–0.86)
	55–59	12	17	11/12 (91.7)	0.76 (0.58–0.93)
		Cases in 1991 (n)	Cases in 2005 (n)		
Age	40–44	38	30	26/38 (68.4)	0.76 (0.65–0.88)
	45–49	39	35	24/39 (61.5)	0.65 (0.52–0.77)
	50–54	59	54	34/59 (57.6)	0.60 (0.49–0.71)
	55–59	12	11	6/12 (50.0)	0.52 (0.27–0.77)
		Cases in 1998 (n)	Cases in 2005 (n)		
Age	40–44	64	57	42/64 (65.6)	0.69 (0.60–0.79)
	45–49	75	66	49/75 (65.3)	0.69 (0.61–0.78)
	50–54	143	131	98/143 (68.5)	0.71 (0.65–0.77)
	55–59	33	28	20/33 (60.6)	0.65 (0.51–0.79)

CI: confidence interval.

$$^a\text{Consistency(\%)} = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100.$$

**Table 7.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes) stratified by BMI.

		Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
		Cases in 1991 (n)	Cases in 1998 (n)		
BMI	Normal weight (<25 kg/m <sup>2</sup> )	62 <sup>b</sup>	62	49/62 (79.0)	0.79 (0.71–0.87)
	Overweight (25–29.9 kg/m <sup>2</sup> )	44 <sup>b</sup>	48	35/44 (79.5)	0.76 (0.66–0.86)
	Obese (≥30 kg/m <sup>2</sup> )	41 <sup>b</sup>	41	29/41 (70.7)	0.70 (0.59–0.82)
		Cases in 1991 (n)	Cases in 2005 (n)		
BMI	Normal weight (<25 kg/m <sup>2</sup> )	62 <sup>b</sup>	59 <sup>c</sup>	44/62 (80.0)	0.73 (0.64–0.82)
	Overweight (25–29.9 kg/m <sup>2</sup> )	44 <sup>b</sup>	35 <sup>c</sup>	21/44 (47.7)	0.53 (0.40–0.66)
	Obese (≥30 kg/m <sup>2</sup> )	41 <sup>b</sup>	35 <sup>c</sup>	25/41 (61.0)	0.65 (0.53–0.78)
		Cases in 1998 (n)	Cases in 2005 (n)		
BMI	Normal weight (<25 kg/m <sup>2</sup> )	99 <sup>d</sup>	89 <sup>e</sup>	74/99 (74.7)	0.79 (0.72–0.85)
	Overweight (25–29.9 kg/m <sup>2</sup> )	99 <sup>d</sup>	83 <sup>e</sup>	59/99 (59.6)	0.65 (0.56–0.73)
	Obese (≥30 kg/m <sup>2</sup> )	114 <sup>d</sup>	106 <sup>e</sup>	74/114 (64.9)	0.79 (0.72–0.85)

CI: confidence interval; BMI: body mass index.

$$^a\text{Consistency(\%)} = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100.$$

<sup>b</sup>The numbers do not add up to 148 due to missing values on height or weight (consequently on BMI).

<sup>c</sup>The numbers do not add up to 130 due to missing values on height or weight (consequently on BMI).

<sup>d</sup>The numbers do not add up to 315 due to missing values on height or weight (consequently on BMI).

<sup>e</sup>The numbers do not add up to 282 due to missing values on height or weight (consequently on BMI).

kappa agreement for type 1 diabetes was weakest in the 1998–2005 test–retest study, which was very close to the kappa agreement for the ~14-year interval in the 1991–2005 test–retest study. In summary, the results show that although the agreement for all self-reported diabetes was weakest in the 1991–2005 test–retest study, this was not the case when analyzing the kappa agreement for the two types of diabetes separately. This suggests that recall problems may not be an

important determinant of the accuracy of self-reported diabetes diagnosis.

One possible reason for the higher kappa agreement among women with type 1 diabetes in our study is that these women may have severe complications sooner<sup>64</sup> than women with type 2 diabetes; this may have contributed the women's recall of age at diagnosis, resulting in a higher agreement for type 1 diabetes.

**Table 8.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes) stratified by physical activity.

		Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
Physical activity level		Cases in 1991 (n)	Cases in 1998 (n)		
	Low	24 <sup>b</sup>	31 <sup>c</sup>	19/24 (79.2)	0.68 (0.54–0.82)
	Medium	106 <sup>b</sup>	101 <sup>c</sup>	80/106 (75.5)	0.77 (0.71–0.84)
	High	11 <sup>b</sup>	11 <sup>c</sup>	8/11 (72.7)	0.73 (0.52–0.94)
Physical activity level		Cases in 1991 (n)	Cases in 2005 (n)		
	Low	24 <sup>b</sup>	27 <sup>d</sup>	18/24 (75.0)	0.74 (0.56–0.85)
	Medium	106 <sup>b</sup>	86 <sup>d</sup>	63/106 (59.4)	0.66 (0.58–0.73)
	High	11 <sup>b</sup>	9 <sup>d</sup>	5/11 (45.5)	0.50 (0.23–0.77)
Physical activity level		Cases in 1998 (n)	Cases in 2005 (n)		
	Low	62 <sup>e</sup>	57 <sup>f</sup>	43/62 (69.4)	0.72 (0.63–0.81)
	Medium	209 <sup>e</sup>	188 <sup>f</sup>	139/209 (66.5)	0.70 (0.65–0.75)
	High	26 <sup>e</sup>	25 <sup>f</sup>	17/26 (65.4)	0.67 (0.52–0.82)

CI: confidence interval.

$$^a\text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100.$$

<sup>b</sup>The numbers do not add up to 148 due to missing values on physical activity level.

<sup>c</sup>The numbers do not add up to 151 due to missing values on physical activity level.

<sup>d</sup>The numbers do not add up to 130 due to missing values on physical activity level.

<sup>e</sup>The numbers do not add up to 315 due to missing values on physical activity level.

<sup>f</sup>The numbers do not add up to 282 due to missing values on physical activity level.

**Table 9.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes) stratified by education level.

		Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
Education level (duration in years)		Cases in 1991 (n)	Cases in 1998 (n)		
	Primary/intermediate (0–9)	35 <sup>b</sup>	40 <sup>c</sup>	28/35 (80)	0.75 (0.64–0.86)
	Secondary (10–12)	63 <sup>b</sup>	70 <sup>c</sup>	50/63 (79.4)	0.75 (0.67–0.83)
	University (13–16)	33 <sup>b</sup>	24 <sup>c</sup>	21/33 (63.6)	0.74 (0.61–0.87)
	Postgraduate and above (17+)	14 <sup>b</sup>	15 <sup>c</sup>	13/14 (92.9)	0.90 (0.78–1.00)
Education level (duration in years)		Cases in 1991 (n)	Cases in 2005 (n)		
	Primary/intermediate (0–9)	35 <sup>b</sup>	32 <sup>d</sup>	18/35 (51.4)	0.54 (0.39–0.68)
	Secondary (10–12)	63 <sup>b</sup>	53 <sup>d</sup>	37/63 (58.7)	0.64 (0.53–0.74)
	University (13–16)	33 <sup>b</sup>	30 <sup>d</sup>	24/33 (72.7)	0.76 (0.64–0.88)
	Postgraduate and above (17+)	14 <sup>b</sup>	12 <sup>d</sup>	10/14 (71.4)	0.77 (0.59–0.95)
Education level (duration in years)		Cases in 1998 (n)	Cases in 2005 (n)		
	Primary/intermediate (0–9)	85 <sup>e</sup>	78 <sup>f</sup>	55/85 (64.7)	0.67 (0.59–0.75)
	Secondary (10–12)	133 <sup>e</sup>	112 <sup>f</sup>	85/133 (64.0)	0.69 (0.62–0.76)
	University (13–16)	63 <sup>e</sup>	61 <sup>f</sup>	45/63 (71.4)	0.72 (0.64–0.81)
	Postgraduate and above (17+)	30 <sup>e</sup>	27 <sup>f</sup>	21/30 (70.0)	0.74 (0.61–0.86)

CI: confidence interval.

$$^a\text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100.$$

<sup>b</sup>The numbers do not add up to 148 due to missing values on education level.

<sup>c</sup>The numbers do not add up to 151 due to missing values on education level.

<sup>d</sup>The numbers do not add up to 130 due to missing values on education level.

<sup>e</sup>The numbers do not add up to 315 due to missing values on education level.

<sup>f</sup>The numbers do not add up to 282 due to missing values on education level.

**Table 10.** Kappa agreements for self-reported diabetes diagnoses (excluding gestational diabetes) stratified by smoking status.

		Test study	Retest study	Consistency (%) <sup>a</sup>	Kappa (95% CI)
		Cases in 1991 (n)	Cases in 1998 (n)		
Smoking status	Never smoker	51	47	38/51 (74.5)	0.78 (0.68–0.87)
	Former smoker	51	50	37/51 (72.5)	0.73 (0.63–0.83)
	Current smoker	46	54	38/46 (82.6)	0.76 (0.67–0.85)
		Cases in 1991 (n)	Cases in 2005 (n)		
Smoking status	Never smoker	51	41	29/51 (56.9)	0.63 (0.51–0.75)
	Former smoker	51	40	31/51 (60.8)	0.68 (0.57–0.79)
	Current smoker	46	49	30/46 (65.2)	0.63 (0.52–0.74)
		Cases in 1998 (n)	Cases in 2005 (n)		
Smoking status	Never smoker	108	94	72/108 (66.7)	0.71 (0.64–0.78)
	Former smoker	103	93	62/103 (60.2)	0.63 (0.55–0.71)
	Current smoker	104	95	75/104 (72.1)	0.75 (0.68–0.82)

CI: confidence interval.

$$^a\text{Consistency}(\%) = \frac{\text{Number of diabetes diagnosis in both test and retest study}}{\text{Number of diabetes diagnosis in test study}} \times 100$$

Since type 2 diabetes typically affects people 40 years of age and over,<sup>51,52</sup> we classified only women aged 40 years and over as having type 2 diabetes. However, it is still possible that women younger than 40 years of age have developed type 2 diabetes.<sup>65–69</sup> In addition, cases identified as having gestational diabetes were excluded from the type 2 diabetes group, although women who had gestational diabetes may develop type 2 diabetes later in life.<sup>70,71</sup> Women aged 39 years or less who reported a diabetes diagnosis (excluding gestational diabetes) were categorized as having type 1 diabetes. Since type 1 diabetes can occur at any age,<sup>72</sup> it is also possible that some of the women classified as having type 2 diabetes in fact had type 1 diabetes. Due to the design and self-reported nature of the study, it was not possible to confirm the exact type(s) of diabetes diagnosis. The results from sensitivity analysis restricting type 1 diabetes cases to those reporting age at diagnosis less than 35 years, and restricting type 2 diabetes to those reporting age at diagnosis more than 44 years, were still acceptable.

This study was larger than previous studies, permitting subgroup analyses. No clear pattern of inconsistency in kappa agreements was observed between different strata of BMI, physical activity, and smoking status. Although no formal test of heterogeneity was performed to assess the statistical difference in kappa agreements across the subgroups, there was a pattern across education groups. The kappa agreement was strongest among the most educated group, while generally it was weaker among the least educated group.

Although the NOWAC cohort is representative of Norwegian women in corresponding age groups, the current sample may not be a representative sample since it includes only the women participating in all the three waves of the study. Furthermore, the respondents with missing values

were excluded. Some research suggests that those belonging to the low socio-economic strata, and are relatively unhealthy, are likely to have a higher proportion of missing values in observational study.<sup>73</sup> Multiple imputation (MI) was not performed, since the kappa statistic<sup>61</sup> is not supported with MI software's<sup>74–77</sup> in Stata. Therefore, the possibility of selection bias limits the external validity of this study.

The kappa agreement we report here is not comparable to other studies<sup>63,78</sup> due to differences in the proportion of people reporting a certain type of diabetes in different studies, or differences in distribution. We found few studies assessing the test–retest reliability of diabetes diagnosis, and the results of those that were found were not consistent. Most showed very good agreement<sup>45–49,79</sup> between the test and the retest studies, while others showed a good<sup>50</sup> or moderate<sup>80</sup> level of agreement. However, most of the studies we found<sup>46–49,80</sup> did not report either the significance probability or the CIs. One possible reason for the higher kappa agreement reported in previous studies<sup>45–50</sup> may be the relatively small time interval between the test and retest studies, as compared to the ~7- or ~14-year interval in our study. The relatively smaller time interval between the test and retest studies may have caused respondents in other populations to remember their previous response more easily, resulting in a higher kappa coefficient.

Another key difference between previous studies<sup>45–50</sup> and our study was their use of interview to collect the information on diabetes diagnosis. As these studies used an interview setting, it is reasonable to assume that the respondent had a chance to ask for questions to be repeated, or for further explanation/clarification, and that the interviewer might have provided it. This may have helped the respondents to understand the question better, and to therefore report more accurately. It is probable that this key difference in



the investigation tool increases the kappa agreement for the test–retest reliability of the studies using interviews to collect data.

However, a study from Manhattan (New York)<sup>80</sup> reported on the test–retest reliability of diabetes diagnosis using telephone interviews. The retest study was conducted within 30 days of the test study, and the kappa agreement between the test and retest studies was found to be 0.48, which is very low considering the short time interval, and despite the use of interviews to collect data. This shows that a short time interval between the test and the retest study and the use of interviews do not necessarily increase the kappa agreement.

The strength of this study is that, it is the first to assess the test–retest reliability of self-reported diabetes diagnosis separately for type 1 and type 2 diabetes. Other strengths of our study include a large cohort size, sensitivity of the estimates by self-reported age at diagnosis, and subgroup analysis within different covariates. This study provides new insights into earlier research by providing the reliability of self-reported diagnosis separately for type 1 and type 2 diabetes.

### Strengths and limitation of this study

- Large (n=33,919) longitudinal population-based study.
- First to assess the test–retest reliability of self-reported diabetes diagnosis separately for type 1 and type 2 diabetes.
- Some women younger than 40 years of age may have developed type 2 diabetes.
- Women with gestational diabetes were excluded, although they may develop type 2 diabetes later in life.

### Conclusion

In conclusion, this study shows that the reliability of the self-reported information on diabetes diagnosis from a large prospective cohort study with long time interval is satisfactory.

### Acknowledgements

The authors are thankful for the anonymous reviewers' comments and suggestions to improve the quality of the article.

### Authors' contributions

This work was completed as part of M.A.S.'s Master's in Public Health thesis, supervised by T.B. M.A.S. performed statistical analyses, data interpretation, and drafted the article. T.B. is the principle investigator of this study. E.L. is the principle investigator of the NOWAC study. E.L. and T.B. critically reviewed the article.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Ethics approval

The NOWAC Study was approved by the Regional Committee for Medical and Health Research Ethics.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### Informed consent

All participants gave written informed consent.

### References

1. Beckett M, Weinstein M, Goldman N, et al. Do health interview surveys yield reliable data on chronic illness among older respondents? *Am J Epidemiol* 2000; 151(3): 315–323.
2. Feinstein AR and Horwitz RI. Double standards, scientific methods, and epidemiologic research. *N Engl J Med* 1982; 307(26): 1611–1617.
3. Goldman N, Lin I-f, Weinstein M, et al. *Evaluating the quality of self-reports of hypertension and diabetes*. Office of Population Research working paper no. 2002-3, <http://westoff.princeton.edu/papers/opr0203.pdf>
4. De Burgos-Lunar C, Salinero-Fort M, Cardenas-Valladolid J, et al. Validation of diabetes mellitus and hypertension diagnosis in computerized medical records in primary health care. *BMC Med Res Methodol* 2011; 11: 146.
5. Robinson JR, Young TK, Roos LL, et al. Estimating the burden of disease: comparing administrative data and self-reports. *Med Care* 1997; 35(9): 932–947.
6. Heliövaara M, Aromaa A, Klaukka T, et al. Reliability and validity of interview data on chronic diseases. The mini-Finland Health Survey. *J Clin Epidemiol* 1993; 46(2): 181–191.
7. Harlow SD and Linet MS. Agreement between questionnaire and medical records. *Am J Epidemiol* 1989; 129(2): 233–248.
8. Huerta JM, José Tormo M, Egea-Caparrós JM, et al. Accuracy of self-reported diabetes, hypertension, and hyperlipidemia in the adult Spanish population. DINO study findings. *Rev Esp Cardiol* 2009; 62(2): 143–152.
9. Midthjell K, Holmen J, Bjørndal A, et al. Is questionnaire information valid in the study of a chronic disease such as diabetes? The Nord-Trøndelag diabetes study. *J Epidemiol Community Health* 1992; 46(5): 537–542.
10. Kriegsman DMW, Penninx BWJH, Van Eijk JTM, et al. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly: a study on the accuracy of patients' self-reports and on determinants of inaccuracy. *J Clin Epidemiol* 1996; 49(12): 1407–1417.
11. Parr C, Veierød M, Laake P, et al. Test-retest reproducibility of a food frequency questionnaire (FFQ) and estimated effects on disease risk in the Norwegian Women and Cancer Study (NOWAC). *Nutr J* 2006; 5(1): 4.
12. Veierød MB, Parr CL, Lund E, et al. Reproducibility of self-reported melanoma risk factors in a large cohort study of Norwegian women. *Melanoma Res* 2008; 18(1): 1–9.
13. Tretli S, Lund-Larsen PG and Foss OP. Reliability of questionnaire information on cardiovascular disease and diabetes:

- cardiovascular disease study in Finnmark county. *J Epidemiol Community Health* 1982; 36(4): 269–273.
14. Jacobsen BK and Bønaa KH. The reproducibility of dietary data from a self-administered questionnaire. The Tromsø study. *Int J Epidemiol* 1990; 19(2): 349–353.
  15. Johansson L, Solvoll K, Opdahl S, et al. Response rates with different distribution methods and reward, and reproducibility of a quantitative food frequency questionnaire. *Eur J Clin Nutr* 1997; 51(6): 346–353.
  16. Hjemdal O, Friberg O, Stiles TC, et al. Resilience predicting psychiatric symptoms: a prospective study of protective factors and their role in adjustment to stressful life events. *Clin Psychol Psychother* 2006; 13(3): 194–201.
  17. Solberg T, Olsen J-A, Ingebrigtsen T, et al. Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. *Eur Spine J* 2005; 14(10): 1000–1007.
  18. Holm I, Friis A, Storheim K, et al. Measuring self-reported functional status and pain in patients with chronic low back pain by postal questionnaires: a reliability study. *Spine* 2003; 28(8): 828–833.
  19. Svege I, Kolle E and Risberg M. Reliability and validity of the Physical Activity Scale for the Elderly (PASE) in patients with hip osteoarthritis. *BMC Musculoskelet Disord* 2012; 13(1): 26.
  20. Sørli T and Sexton HC. The factor structure of “The Ways of Coping Questionnaire” and the process of coping in surgical patients. *Pers Individ Differ* 2001; 30(6): 961–975.
  21. Bjertnaes O, Iversen H and Kjøllesdal J. PIPEQ-OS—an instrument for on-site measurements of the experiences of inpatients at psychiatric institutions. *BMC Psychiatr* 2015; 15: 234.
  22. Nordberg SS, Moltu C and Råbu M. Norwegian translation and validation of a routine outcomes monitoring measure: the treatment outcome package. *Nordic Psychology*. Epub ahead of print 15 September 2015. DOI: 10.1080/19012276.2015.1071204.
  23. Bjørnå HB, Hillesund ER, Torstveit MK, et al. An assessment of the test-retest reliability of the New Nordic Diet score. *Food Nutr Res* 2015; 59: 28397.
  24. Moljord IEO, Lara-Cabrera ML, Perestelo-Pérez L, et al. Psychometric properties of the Patient Activation Measure-13 among out-patients waiting for mental health treatment: a validation study in Norway. *Patient Educ Couns* 2015; 98(11): 1410–1417.
  25. Myr R, Bere E and Overby N. Test-retest reliability of a new questionnaire on the diet and eating behavior of one year old children. *BMC Res Notes* 2015; 8(1): 16.
  26. Røysamb E, Vittersø J and Tambs K. The relationship satisfaction scale—psychometric properties. *Norsk Epidemiologi* 2014; 24(1–2): 187–194.
  27. Øverby NC, Hillesund ER, Sagedal LR, et al. The Fit for Delivery study: rationale for the recommendations and test-retest reliability of a dietary score measuring adherence to 10 specific recommendations for prevention of excessive weight gain during pregnancy. *Matern Child Nutr* 2015; 11(1): 20–32.
  28. Bjertnaes O, Skudal KE, Iversen HH, et al. The Patient-Reported Incident in Hospital Instrument (PRIH-I): assessments of data quality, test-retest reliability and hospital-level reliability. *BMJ Qual Saf* 2013; 22(9): 743–751.
  29. Johansen JB, Roe C, Bakke E, et al. Reliability and responsiveness of the Norwegian version of the Neck Disability Index. *Scand J Pain* 2014; 5(1): 28–33.
  30. Nordtorp HL, Nyquist A, Jahnsen R, et al. Reliability of the Norwegian Version of the Children’s Assessment of Participation and Enjoyment (CAPE) and Preferences for Activities of Children (PAC). *Phys Occup Ther Pediatr* 2013; 33(2): 199–212.
  31. Løchting I, Grotle M, Storheim K, et al. Individualized quality of life in patients with low back pain: reliability and validity of the Patient Generated Index. *J Rehabil Med* 2014; 46(8): 781–787.
  32. Iversen M, Espehaug B, Rokne B, et al. Psychometric properties of the Norwegian version of the Audit of Diabetes-Dependent Quality of Life. *Qual Life Res* 2013; 22(10): 2809–2812.
  33. Haldorsen B, Svege I, Roe Y, et al. Reliability and validity of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire in patients with shoulder impingement syndrome. *BMC Musculoskelet Disord* 2014; 15: 78.
  34. Aaby C and Heimdal J-H. The voice-related quality of life (V-RQOL) measure—a study on validity and reliability of the Norwegian version. *J Voice* 2013; 27(2): 258.e29–258.e33.
  35. Tavoly M, Jelsness-Jørgensen L-P, Wik H, et al. Quality of life after pulmonary embolism: first cross-cultural evaluation of the pulmonary embolism quality-of-life (PEmb-QoL) questionnaire in a Norwegian cohort. *Qual Life Res* 2015; 24(2): 417–425.
  36. Kapstad H, Nelson M, Øverås M, et al. Validation of the Norwegian short version of the Body Shape Questionnaire (BSQ-14). *Nord J Psychiatry* 2015; 69(7): 509–514.
  37. Klokkeud M, Grotle M, Løchting I, et al. Psychometric properties of the Norwegian version of the patient generated index in patients with rheumatic diseases participating in rehabilitation or self-management programmes. *Rheumatology* 2013; 52: 924–932.
  38. Amble I, Gude T, Stubdal S, et al. Psychometric properties of the outcome questionnaire-45.2: the Norwegian version in an international context. *Psychother Res* 2014; 24(4): 504–513.
  39. Agerup T, Lydersen S, Wallander J, et al. Maternal and paternal psychosocial risk factors for clinical depression in a Norwegian community sample of adolescents. *Nord J Psychiatry* 2015; 69(1): 35–41.
  40. Skre I, Friberg O, Elgaroy S, et al. The factor structure and psychometric properties of the Clinical Outcomes in Routine Evaluation—Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatr* 2013; 13: 99.
  41. Erdvik IB, Øverby NC and Haugen T. Translating, reliability testing, and validating a Norwegian Questionnaire to Assess Adolescents’ Intentions to be Physically Active After High School Graduation. *SAGE Open*. Epub ahead of print 13 April 2015. DOI: 10.1177/2158244015580374.
  42. Bergland Å, Hofoss D, Kirkevold M, et al. Person-centred ward climate as experienced by mentally lucid residents in long-term care facilities. *J Clin Nurs* 2015; 24(3–4): 406–414.
  43. Østerås N, Garratt A, Grotle M, et al. Patient-reported quality of care for osteoarthritis: development and testing of the osteoarthritis quality indicator questionnaire. *Arthritis Care Res* 2013; 65(7): 1043–1051.
  44. Trochim WMK. Types of reliability, <http://www.social-researchmethods.net/kb/relytypes.php> (2006, accessed 24 December 2011).
  45. Andresen EM, Malmstrom TK, Miller DK, et al. Retest reliability of self-reported function, self-care, and disease history. *Med Care* 2005; 43(1): 93–97.

46. Yount BW, Wyrwich KW and Brownson RC. The reliability of a questionnaire-based metabolic syndrome surveillance tool. *Metab Syndr Relat Disord* 2007; 5(3): 282–289.
47. Brownson RC, Jackson-Thompson J, Wilkerson JC, et al. Reliability of information on chronic disease risk factors collected in the Missouri Behavioral Risk Factor Surveillance System. *Epidemiology* 1994; 5(5): 545–549.
48. Starr GJ, Grande ED, Taylor AW, et al. Reliability of self-reported behavioural health risk factors in a South Australian telephone survey. *Australian N Z J Public Health* 1999; 23(5): 528–530.
49. Bosetti C, Tavani A, Negri E, et al. Reliability of data on medical conditions, menstrual and reproductive history provided by hospital controls. *J Clin Epidemiol* 2001; 54(9): 902–906.
50. Bowlin SJ, Morrill BD, Nafziger AN, et al. Reliability and changes in validity of self-reported cardiovascular disease risk factors using dual response: the behavioral risk factor survey. *J Clin Epidemiol* 1996; 49(5): 511–517.
51. Khardori R. Type 2 diabetes mellitus, <http://emedicine.medscape.com/article/117853-overview> (2011, accessed 24 December 2011).
52. Sheikh MA, Lund E and Braaten T. The predictive effect of body mass index on type 2 diabetes in the Norwegian women and cancer study. *Lipids Health Dis* 2014; 13: 164.
53. Lund E. The Norwegian Women and Cancer study, NOWAC, <http://site.uit.no/nowac/>; <http://uit.no/> (accessed 14 October 2011).
54. The Norwegian Women and Cancer study (NOWAC). *The Norwegian Women and Cancer study*. Tromsø: Department of Community medicine, Faculty of medicine, The University of Tromsø, 2008 (updated 4 July 2008), <http://site.uit.no/nowac/> (accessed 21 December 2015).
55. Lund E, Dumeaux V, Braaten T, et al. Cohort profile: the Norwegian Women and Cancer Study—NOWAC—Kvinner og kreft. *Int J Epidemiol* 2008; 37(1): 36–41.
56. Lund E, Kumle M, Braaten T, et al. External validity in a population-based national prospective study—the Norwegian Women and Cancer Study NOWAC. *Cancer Causes Control* 2003; 14(10): 1001–1008.
57. Skeie G, Mode N, Henningsen M, et al. Validity of self-reported body mass index among middle-aged participants in the Norwegian Women and Cancer study. *Clin Epidemiol* 2015; 7: 313–323.
58. Hjartåker A, Andersen LF and Lund E. Comparison of diet measures from a food-frequency questionnaire with measures from repeated 24-hour dietary recalls. The Norwegian Women and Cancer Study. *Public Health Nutr* 2007; 10(10): 1094–1103.
59. Brustad M, Skeie G, Braaten T, et al. Comparison of telephone vs face-to-face interviews in the assessment of dietary intake by the 24h recall EPIC SOFT program—the Norwegian calibration study. *Eur J Clin Nutr* 2003; 57(1): 107–113.
60. Fleiss JL, Levin B and Paik MC. *Statistical methods for rates and proportions* (ed WA Shewart and SS Wilks). 3rd ed. Hoboken, NJ: John Wiley & Sons, 2003, 800 pp.
61. Reichenheim ME. Confidence intervals for the kappa statistic. *Stata J* 2004; 4(4): 421–428.
62. Simon S. *What is a kappa coefficient?* (Cohen's kappa). Kansas City, MO: Children's Mercy Hospitals and Clinics, 2008 (updated 14 July 2008), <http://www.childrensmc.org/stats/definitions/kappa.htm> (accessed 24 December 2011).
63. StatsDirect Limited. *Kappa and Maxwell*. Altrincham: StatsDirect Limited, 2011. (updated 2011), <http://www.ukph.org/help/stats-direct.htm#agreement/kappa.htm> (accessed 24 December 2011).
64. Lab Tests Online. Diabetes, <http://labtestsonline.org/understanding/conditions/diabetes?start=1> (accessed 24 December 2011).
65. World Health Organization (WHO). *Diabetes*. Geneva: WHO, 2011 (updated August 2011), <http://www.who.int/mediacentre/factsheets/fs312/en/index.html> (accessed 23 December 2011).
66. Chyun DA, Wackers FJ, Inzucchi SE, et al. Autonomic dysfunction independently predicts poor cardiovascular outcomes in asymptomatic individuals with type 2 diabetes in the DIAD study. *SAGE Open Med*. Epub ahead of print 24 February 2015. DOI: 10.1177/2050312114568476.
67. AL-Aboudi IS, Hassali MA, Shafie AA, et al. A cross-sectional assessment of health-related quality of life among type 2 diabetes patients in Riyadh, Saudi Arabia. *SAGE Open Med*. Epub ahead of print 9 October 2015. DOI: 10.1177/2050312115610129.
68. Shaw J. Epidemiology of childhood type 2 diabetes and obesity. *Pediatr Diabetes* 2007; 8: 7–15.
69. Jenssen TG, Tonstad S, Claudi T, et al. The gap between guidelines and practice in the treatment of type 2 diabetes: a nationwide survey in Norway. *Diabetes Res Clin Pract* 2008; 80(2): 314–320.
70. World Health Organization (WHO) Regional Office for Europe. *Facts and figures Copenhagen*, <http://www.euro.who.int/en/what-we-do/health-topics/noncommunicable-diseases/diabetes/facts-and-figures> (2011, accessed 24 December 2011).
71. Ben-Haroush A, Yogev Y and Hod M. Epidemiology of gestational diabetes mellitus and its association with Type 2 diabetes. *Diabet Med* 2004; 21(2): 103–113.
72. Nasjonalt folkehelseinstituttet (Norwegian Institute of Public Health). Forekomsten av type 1-diabetes har økt med omlag 30 prosent 2008, [http://www.fhi.no/eway/default.aspx?pid=233&trg=MainLeft\\_5565&MainArea\\_5661=5565:0:15,1212:1:0:0:::0:0&MainLeft\\_5565=5544:64250::1:5569:1:::0:0](http://www.fhi.no/eway/default.aspx?pid=233&trg=MainLeft_5565&MainArea_5661=5565:0:15,1212:1:0:0:::0:0&MainLeft_5565=5544:64250::1:5569:1:::0:0) (accessed 24 December 2011).
73. Sheikh MA, Abelsen B and Olsen JA. Role of respondents' education as a mediator and moderator in the association between childhood socio-economic status and later health and wellbeing. *BMC Public Health* 2014; 14(1): 1172.
74. Royston P and White IR. Multiple Imputation by Chained Equations (MICE): implementation in Stata. *J Stat Software* 2011; 45(4): 1–20.
75. White IR, Royston P and Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30(4): 377–399.
76. Carpenter JR and Kenward MG. *Multiple imputation and its application*. Sussex: John Wiley & Sons, 2013, 364 pp.
77. StataCorp. *Stata 14 multiple-imputation reference manual*. College Station, TX: Stata Press, 2015, <http://www.stata.com/manuals14/mi.pdf>
78. Uebersax J. Kappa coefficients: John Uebersax Enterprises LLC (updated 18 March 2010), <http://www.john-uebersax.com/stat/kappa.htm> (2010, accessed 24 December 2011).
79. Schneider ALC, Pankow JS, Heiss G, et al. Validity and reliability of self-reported diabetes in the Atherosclerosis Risk in Communities Study. *Am J Epidemiol* 2012; 176(8): 738–743.
80. Kargman DE, Sacco RL, Boden-Albala B, et al. Validity of telephone interview data for vascular disease risk factors in a racially mixed urban community: the Northern Manhattan Stroke Study. *Neuroepidemiology* 1999; 18(4): 174–184.