



Longitudinal Diabetic Status Dataset

Baseline through Visit 17

CODEBOOK

ARCHIVED DATASET 2026

Dataset file niadiab2019

PAGE INTENTIONALLY BLANK

DOCUMENTATION FOR THE SWAN LONGITUDINAL DIABETIC STATUS DATASET

The Diabetes Status dataset is a longitudinal dataset (multiple lines of data per participant) of the diabetic status in all SWAN participants from Baseline to Visit 17. The assigned participant ID has been replaced with a randomly generated ARCHID in order to protect participant privacy. A variable describing the race/ethnicity of participants (RACE) was added from the Screener dataset.

Differences from the prior release of the Diabetic Status dataset (niadiab2026):

- Visits 17 was added to the dataset.

Who is included in the public use dataset:

The dataset contains information for the 3,302 women from the seven clinical sites participating in the SWAN longitudinal study. The sites include 11, 12, 13, 14, 15, 16 and 17.

Criteria used to create the diabetes variable (DIAB):

Diabetes status = Yes (DIAB = 1):

1. Prevalent case:
 - a. Must have more than one time point for classification (exclude women with a baseline only diabetes assignment when their only participation is at baseline)
 - b. Women with more than one time point for classification will have the following possible assignments:
 - i. Prevalent diabetes
 - ii. Prevalence of a high probability of diabetes
 - iii. Prevalence of a low probability of diabetes
2. Incident case
 - a. Must have more than one time point for classification
 - b. Must have baseline that is classified as being non-diabetic
 - c. For survival analysis, assignment to one of three incident states will occur the first visit in which a diabetes flag is raised. The three incident states are:
 - i. Diabetes
 - ii. High probability of diabetes
 - iii. Lower probability of diabetes

Case definitions are based on patterns:

- a. Frequency: more than 70% of the visits have a diabetes flag, particularly after the first “diabetes” flag is raised
- b. Consistency: 3 or more consecutive visits with a diabetes flag
- c. Temporality: onset of pattern suggests presentation of frequency and consistency in progression to later visits.

Diabetes status = No (DIAB = 0):

Participant recorded as 0 (“No”) when:

- The woman had no self-reported diabetes
- The woman’s glucose results < 126
- The woman never took anti-diabetic medications

If all the three criteria were satisfied for a specific visit, or 2 criteria were satisfied and the third one was missing, then “No” was assigned as the woman’s diabetic status for that visit (DIAB = 0).

Diabetes status = Lower Probability of Diabetes (DIAB = 2):

Participant recorded as 2, “Lower Probability of Diabetes” when there is evidence of diabetes at some visits (usually less than 3) but does not meet criteria.

Diabetes status = Unknown (DIAB = .C):

Diabetes status was assigned as .C in three situations:

- When data on all the three criteria were available and any one of the three criteria was violated then “Unknown” was assigned as the women’s diabetic status (DIAB = .C).

- When data on one criterion of the above specified criteria were missing and at least one of the available two criteria was violated then “Unknown” was assigned as the women’s diabetic status (DIAB=.C).
- When data on two criteria of the above specified criteria were missing, then “Unknown” was assigned as the women’s diabetic status (DIAB = .C).

Reclassification of “Unknown” status:

1. Women were reclassified from being “Unknown” to “diabetic” at specific time points if their glucose level was ≥ 126 while not be on steroid for two consecutive visits AND /OR they reported use of antidiabetic medications (DIAB = 1)
2. Women were reclassified from being “Unknown” to “low probability diabetic” (DIAB = 2) if:
 - Their glucose level was ≥ 126 (they should not be on steroid) at only one time point
 - OR If they only self-reported diabetes
3. Women who did not meet any of the above remained as “unknown”

Diabetes status = MISSING (DIAB = .A):

Missing was assigned only when data on all the above 3 criteria were missing (i.e. we do not have data on any of the three criteria).

Flag Variable (IMPFLG):

IMPFLG is a created variable to flag visits where the created DIAB variable values based on the above strategies were modified from ‘.C’ or ‘.A’ to “NO”.

We replaced ‘.C’ and ‘.A’ with “NO” as the woman’s diabetic status for a visit when the ‘.C’ or ‘.A’ visit was followed by at least one visit with DIAB = NO and DIAB status was “NO” in all available visits before that visit.

IMPFLG = 1 if ‘.A’ was replaced with “NO”; IMPFLG = 2 if ‘.C’ was replaced with “NO”.

Recommendation for using this dataset:

For diabetes related analyses as well as any risk score calculation that will require classifying participants based on their diabetes status, investigators should only use DIAB = 1 as an indication for confirmed diabetic cases. Both DIAB = 2 and .C should be set to missing.

Missing data coding:

Original missing codes (-1: not applicable, -7: refused, -8: don’t know, -9: missing) have been recoded to SAS missing codes (.B: not applicable, .D: refused, .C: don’t know, and .A: missing).

NOTE: Please be aware that there was for the site 16 women a gap of 2-3 visits between Visits 06 and 09. This might affect the accuracy of the created variables.

Variables included in the dataset:

Variable	Label	Codes
ARCHID	Participant ID	
VISIT	Study visit	"00" = Baseline to "17" = Visit 17
RACE	Race/Ethnicity	1 = Black 2 = Chinese/Chinese American 3 = Japanese/Japanese American 4 = White Non-Hispanic 5 = Hispanic
DIAB	Diabetic Status	0 = No 1 = Diabetic 2 = Lower Probability of Diabetes .A = Missing .C = Unknown
DIABVISIT	Diabetes Onset Visit	"00" = Baseline to "17" = Visit 17
IMPFLG	Imputation flag	0 = No Imputation for DIAB variable 1 = .A (Missing) changed to "No" for DIAB variable 2 = .C (Unknown) changed to "No" for DIAB variable