IOE, Faculty of Education and Society

Harmonised indicators of self-reported diabetes in five British cohort studies

User Guide (Version 1)

June 2025

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Laura Gimeno, Rebecca Hardy, Martina Narayanan.

How to cite this guide

Gimeno L., Hardy M., Narayanan M. (2025) *Harmonised Indicators of Self-Reported Diabetes in Five British Cohort Studies (Version 1)*. London: UCL Centre for Longitudinal Studies.

Data citation and acknowledgement

You should cite the data and acknowledge CLS following the guidance from cls.ucl.ac.uk/data-access-training/citing-our-data/

Centre for Longitudinal Studies

Centre for Longitudinal Studies (CLS)

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre. It is home to a unique series of UK national cohort studies. It is part of the UCL Social Research Institute, based at the IOE, UCL's Faculty of Education and Society.

This document is available in alternative formats. Please email the Centre for Longitudinal Studies at clsdata@uck.ac.uk

Contents

Acknowledgements

This project was supported by the Economic and Social Research Council (ESRC) which finds the UCL Centre for Longitudinal Studies (grant numbers ES/M001660/1 and ES/W013142/1). The ESRC had no role in the design, execution, analysis, or interpretation of the data or in the writing up of the findings of this report.

This project brings together data from five British cohorts: the Medical Research Council (MRC) National Survey of Health and Development (NSHD), the 1958 National Child Development Survey (NCDS), the 1970 British Cohort Study (BCS), Next Steps (formerly the Longitudinal Study of Young People in England), and the Millennium Cohort Study (MCS). NSHD is funded by the MRC and is hosted by the MRC Unit for Lifelong Health and Ageing at UCL. NCDS, BCS70, Next Steps, and MCS receive core funding from the ESRC and are hosted by the Centre for Longitudinal Studies at UCL. The most recent sweep of BCS70, at age 46, received additional funding from the MRC and the British Heart Foundation.

The authors would like to thank the owners of the five studies included in this report, and the cohort members and their families who have given their time to take part in these studies. We would also like to acknowledge the UK Data Service for providing access to NCDS, BCS70, Next Steps, and MCS, and the MRC Lifelong Health and Ageing Unit for access to NSHD.

1. Introduction

The British cohort studies and other longitudinal populations studies collect information on health conditions and diagnoses from self-reports at interviews and on questionnaires.

Researchers may be interested in variables which indicate presence of a health condition for several reasons. They may be interested in a health condition as an outcome in analyses. The cohort studies may be particularly valuable for investigating early life factors in relation to health outcomes given the wealth of prospectively collected data. Alternatively, health conditions may be used as an independent variable in analyses, as a confounder, or to exclude participants with existing health conditions from analyses.

Diabetes is an increasingly common health-related condition in the UK (Holden *et al.,* 2009), and recent estimates for 2021-2022 suggest that there are 4.3 million people in the UK living with diagnosed diabetes (Diabetes UK, n.d.). Diabetes is linked to a range of serious complications including stroke and cardiovascular disease and can impact quality of life and mental health. Combining data from multiple longitudinal studies can be used to increase the power of analyses with diabetes as the outcome, and multiple cohorts can be used to compare diabetes experience across generations. Access to harmonised data on diabetes may therefore be a valuable resource. To carry out cross-cohort analyses appropriately, careful consideration of how comparable the health measures are within studies over time and between studies is required (Bann et al., 2022). In this report we describe the process of harmonising one such condition: diabetes mellitus.

1.1 Aims

To document all self-reported measures related to diabetes in five British cohort studies, and to produce harmonised indicators of self-reported diabetes.

1.2 Cohorts included

This resource considers information on self-reported diabetes in the following five studies: the MRC National Survey of Health and Development (NSHD), the 1958 National Child Development Study (NCDS), the 1970 British Cohort Study (BCS70), Next Steps, and the Millennium Cohort Study.

The MRC National Survey of Health and Development. The MRC NSHD is Britain's longest running birth cohort study. It originally consisted of a social stratified sample (N=5,362) of singleton births to married parents in England, Scotland, and Wales in March 1946 (Wadsworth *et al.,* 2006). The sample was selected from an initial maternity survey of 13,637 pregnancies and consisted of all births to fathers in non-manual and agricultural occupations, and a random 1-in-4 sample of births to fathers in manual occupations. To date, the participants have been followed 29 times between ages 2 and 77 years, including three COVID-19 surveys in 2020 and 2021. More information is available on the MRC LHA website.

The 1958 National Child Development Study. The NCDS follows the lives of 17,415 people born in England, Scotland, or Wales in a single week in 1958 (Power & Elliott, 2006). The NCDS began in 1958 with the Perinatal Mortality Survey, which captured 98% of the total births in Great Britain in the target week. The cohort has been followed a total of 10 times between ages 7 and 55 years (including a biomedical survey in 2002, and three COVID-19 surveys in 2020 and 2021). More information is available on the CLS website.

The 1970 British Cohort Study. BCS70 follows the lives of 17,198 people born in England, Scotland, and Wales in a single week of 1970 (Elliott & Shepherd, 2006; Sullivan *et al.*, 2023). BCS70 began as the British Births Survey, and participants have since been followed up 10 times to date, between ages 5 to 51 years (including a biomedical sweep in 2016-2018, and three COVID-19 surveys in 2020 and 2021). More information is available on the CLS website.

Next Steps. Next Steps (formerly known as the Longitudinal Study of Young People in England) follows a sample of over 16,000 people born in 1989/90. Cohort members were recruited through schools in England when they were aged 13-14 years in 2004. Cohort members were interviewed annually between ages 14 and 20

years, at age 25 years, in three COVID-19 Survey sweeps in 2020 and 2021 and at age 32. More information is available on the CLS website.

The Millennium Cohort Study. MCS follows the lives of 19,517 children born in England, Scotland, Wales, and Northern Ireland in 2000-2002 (Connelly & Platt, 2014). Since the initial birth survey at 9 months, there have been six more follow-up surveys at ages 3, 5, 7, 11, 14, and 17 years, as well as an additional three COVID-19 survey sweeps. More information is available on the CLS website.

1.3 Question types used

In this version of harmonised indicators of diabetes across the cohorts, we use safeguarded survey data for NCDS, BCS70, Next Steps, and MCS, which are freely available from the UK Data Service (UKDS), subject to the UKDS End User Licence. NSHD data were obtained through Skylark.

The derivation of the harmonised indicators is primarily based on self-reported data. We additionally leverage some information from doctors' reports in childhood sweeps which summarise information from parents' reports and medical records. Biomarkers and linked health data were not included. We focus on measures that were administered to entire cohorts only (i.e., only information from questionnaires aimed at the entire cohort were considered). We focus only on information on the diabetes status of cohort members themselves (not their parents or children).

All datasets used in the harmonisation project are listed in appendix 1.

2. Retrospective harmonisation of diabetes

Retrospective harmonisation is a term used to describe the process of manipulating data within or across existing studies with the aim of making them more comparable. While each harmonisation project is unique, and decisions must be made on a case-by-case basis depending on the data in question, there are broad methodological guidelines available for harmonisation (Fortier *et al.,* 2017), which we have aimed to follow.

2.1 Documentation of all diabetes questions

The first stage of the harmonisation process involved examining all questionnaires administered to whole cohorts across the five cohort studies. For each questionnaire, questions on diabetes, and questions on chronic health conditions for which it would be possible to report diabetes in some way (for example, for longstanding illness questions where outcomes were provided in a fine-grained enough format for cases of diabetes to be identified) were documented. Specific attention was paid to the order and routing of questions. Other self-reported questions related to diabetes include variables such as age at first diagnosis, medication (including insulin use), and visits to the doctor. A list of questions on diabetes (or susceptible to contain some information on diabetes) can be found in on https://github.com/CLS-Data/harmonised-diabetes-across-cohorts.

All variables used in the derivation of the harmonised diabetes variables are listed in appendix 2.

2.2 Creation of harmonised indicator variables for diabetes

2.2.1 Sweep-specific indicators

Three main types of question related to diabetes were identified to be consistently available across different sweeps and across cohorts, and therefore most suitable to

feed into the creation of harmonised indicator variables. They capture either lifetime ('ever') prevalence or point ('current') prevalence.

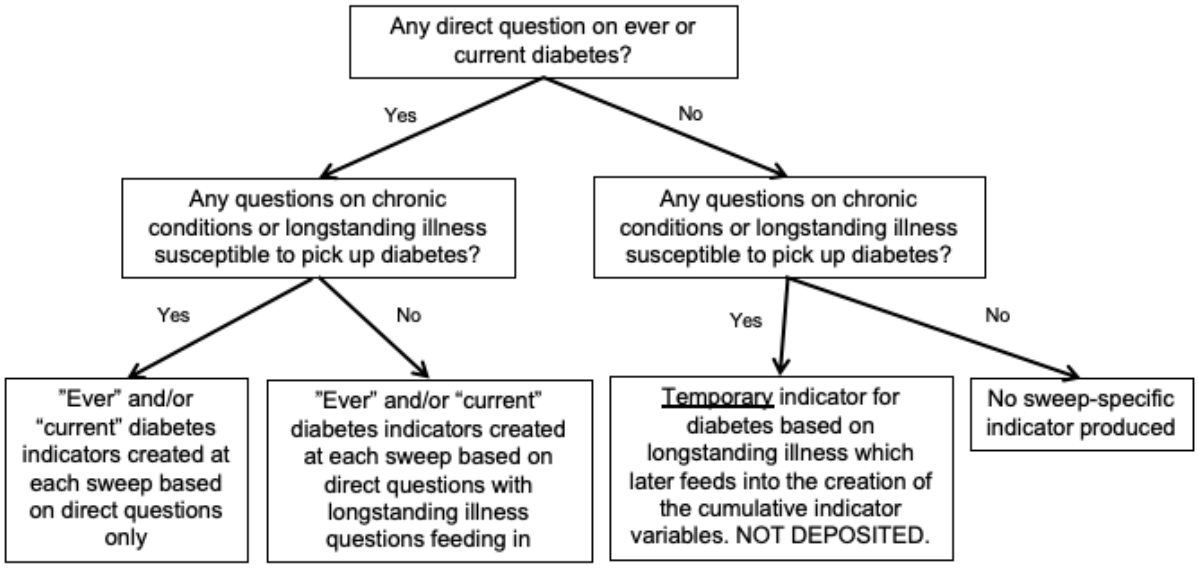
- Direct questions about lifetime prevalence of diabetes. These questions ask whether the cohort member has ever had diabetes, or whether they have had diabetes since the previous sweep (in which case these responses can be combined with those from previous sweeps to derive lifetime prevalence).
- Direct questions about point prevalence of diabetes. These questions ask cohort members if they currently have diabetes at the time of the sweep, or in the last 12 months.
- Other questions susceptible to provide additional information on point prevalence of diabetes. These include questions about any longstanding illnesses where a cohort member reports a condition which is then coded using the International Classification of Diseases (ICD).

Information from questions which asked directly about whether the participant had diabetes were prioritised. Questions on longstanding illness and other questions likely to pick up reports of chronic conditions were used to feed into the within-sweep indicator variables created based on these direct questions. If such direct questions were not available in a particular sweep, questions on longstanding illness alone were used to create intermediate variables which later fed into the creation of cumulative indicator variables (**Figure 1**). We did not include within-sweep indicators based on longstanding illness type questions alone in the deposited dataset because the way these questions were asked leads to more selective response rates and thus biased prevalences (to be able to provide an ICD-10 code for diabetes, respondents first need to self-report a longstanding illness, which not all respondents with diabetes do). In **Box 1**, we show a worked example of how most sweep-specific indicators in the deposited dataset were derived.

Indicators of prevalence (based on whether ever had or currently have diabetes) were derived for each sweep. These variables do not use information from other sweeps and include only those who provided information at each given sweep. Exceptions to this are some of the "ever" indicators in the 1946 and 1970 cohort studies, which were cumulated across sweeps because of how questions were asked. In the 1946 cohort, information on previous reports of diabetes were fed

forward, so cohort members with a previous report of diabetes were not asked about their diabetes status again. In the 1970 cohort, participants were sometimes asked whether they "[had] ever had diabetes *since the last sweep*?" rather than if they "[had] ever had diabetes?" Study response files were used to ensure that missing data was coded consistently, to enable differentiation between item and sweep non-response.

Figure 1. Questions feeding into sweep-specific indicators of diabetes.



Box 1. Worked example of derivation process at age 34 in BCS70

There is a direct question on diabetes prevalence in this sweep.

Since [last interview/January 2000] have you had any of the health problems listed on this card? (Sugar) diabetes.

Later in the questionnaire, there is a question which could indirectly capture point prevalence:

Do you have any longstanding illness, disability, or infirmity? What [else] is the matter with you?

Here, we create a sweep-specific indicator that is enriched with responses to the longstanding illness question. If a cohort member reported "Yes" and had an ICD-10 code indicative of diabetes (E10 to E14), we forced the sweep indicator to take a "Yes" value.

2.2.2. Cross-sweep indicators

Cumulative indicators

Since questions on diabetes in the two youngest cohorts (Next Steps and MCS) were restricted to information collected during the COVID-19 pandemic, cross-sweep harmonisation focused on the three oldest cohorts (born in 1946, 1958 and 1970). We created cumulative 'ever' reported diabetes indicators including all sweep where questions on diabetes had been asked. Multiple cumulative indicator variables were created reflecting whether a cohort member had ever reported diabetes up to a certain time-point. For each sweep, if a cohort member answered 'yes' to an ever or current diabetes question or had a code indicative of diabetes in response to a longstanding illness-type question at that or previous sweeps, the indicator takes a value of 1 (yes). If the respondent had not ever reported diabetes up to that point and had a report of never diabetes at this or subsequent sweeps, the indicator takes a value of 0 (no). A small number of cohort members were left without a determined diabetes status at each sweep. A larger number of cohort members have missing values on the cumulative indicator due to sweep non-response.

Diabetes type indicators

To derive diabetes type, we leveraged information on self-reported type, age at diagnosis, insulin use, and information on the sweep at which cohort members reported diabetes for the first time (**Figure 2**). We assigned cohort members based on their self-reported type in the first instance. For those without a self-reported type, we then worked through an algorithm to assign a type using other self-reported information.

We first identified likely cases of gestational diabetes in women by identifying cases with a self-reported date of diagnosis within one year of the birth of a child, where women didn't subsequently report current diabetes or insulin use. Gestational diabetes develops during pregnancy but usually goes away after birth. However, research suggests that women who have had gestational diabetes are at far higher risk of developing type 2 diabetes (Vounzoulaki *et al.*, 2020). In cases were a woman thought to have initially had gestational diabetes had reports of current diabetes and

insulin in adulthood consistent with type 2 diabetes, they were classified as having type 2 diabetes rather than gestational diabetes.

We then leveraged information on insulin as different types of diabetes likely have different patterns of insulin use. Type 1 diabetes is a condition where the body cannot produce insulin, and those with type 1 diabetes are treated with insulin from the time they are diagnosed. Type 2 diabetes, in contrast, can be managed and treated in a variety of ways, including insulin use. Not using insulin or not beginning insulin use immediately is therefore an indicator of *not* having type 1 diabetes. However, additional cases of type 1 diabetes could not be identified using insulin, since some individuals with type 2 diabetes may also be prescribed insulin as soon as their diabetes has been diagnosed. Those who reported not using insulin or who began using insulin more than two years after the reported date of diabetes diagnosis were assigned to the type 2 group.

We then used age at diagnosis. Those with a self-reported age of onset <20 years were assigned type 1, whereas those with an age of onset >40 years were assigned type 2. While type 1 diabetes if often diagnosed in childhood, type 2 diabetes is generally diagnosed in adulthood. If age at diagnosis was not available but age when the cohort member began injecting insulin was, then we used the latter to allocate type, with the assumption that among those with type 1 diabetes, this would be very close to age of diagnosis, whereas in those with type 2 diabetes, insulin use would typically begin several years after diabetes diagnosis.

We used patterns of response to the sweep-specific current and ever indicators created in the first step of harmonisation on a case-by-case basis to assign type for a small number of additional cases. If cohort members had responded “Yes” to currently having diabetes at older sweeps (in their fifties and beyond) and there was some evidence of not having had diabetes in their twenties or thirties, we assigned cases to the type 2 category. If there was evidence for diabetes in childhood sweeps, we assigned cases to type 1 diabetes.

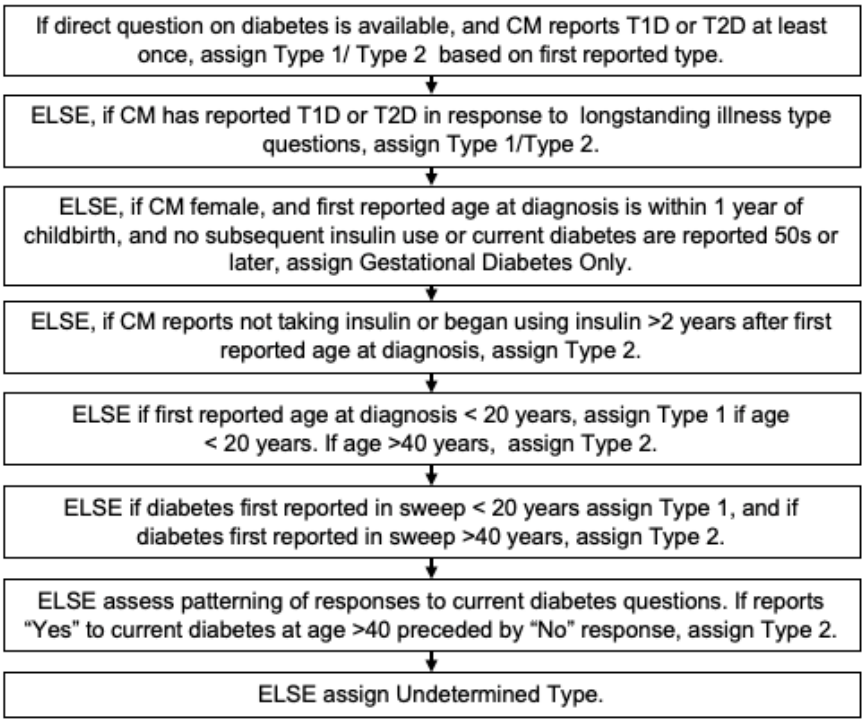
Even after this process, a small number of cases remained unallocated. These were typically cases where no information on diabetes was available in childhood or early adulthood to distinguish between type 1 diabetes that a cohort member may have had since childhood and acquired type 2 diabetes. Other remaining cases are those

with inconsistent patterns of response. Since a rule to allocate these cases could not be defined, these were left in a separate group.

We created an indicator which holds information on the reason that a particular cohort member was assigned type 1, type 2, or gestational diabetes.

We also created flags for cases where age at diagnosis and reported insulin use were inconsistent with self-reported type. In some instances, cohort members self-reported having type 1 diabetes but also reported not using insulin, or an age at diagnosis >30 years, or that they had type 2 diabetes but reported an age at diagnosis <20 years. An accompanying variable contains information on why the flag was raised for each case.

Figure 2. Ascertaining diabetes type based on self-report.



2.2.3 Cross-study harmonisation

The process of cross-study harmonisation started with selecting variables related to diabetes that were available across cohorts and that were worded in a similar enough way to assume that they capture the same construct. We then used the same process to derive harmonised variables in each cohort.

2.3 Quality of the harmonised variables

2.3.1 Agreement between self-reported diabetes and other measures

Previous work has demonstrated that self-reported diabetes shows good agreement with GP records in NSHD (Pastorino *et al.*, 2015), and there is good agreement in NCDS and BCS70 between self-reported diabetes and measures of glycated haemoglobin (HbA1c) taken in biomedical sweeps accounting for medication use (Gondek, 2020). That is, there are few false positive cases, whereby those who report having diabetes are usually confirmed as such using measures other than self-report.

Measures based on self-reported diabetes miss undiagnosed cases of diabetes. It is, however, useful to separate diagnosed cases separately from undiagnosed cases since diagnosis could lead to positive changes in lifestyle behaviour, although apart from giving up smoking longitudinal studies have found little evidence to support this hypothesis (Hackett *et al.,* 2018). It is also possible that the consistently observed relationship between diabetes and depression is, at least in part, due to the burden of living with a diabetes diagnosis and the associated treatment (Chen *et al.,* 2016). It should also be noted that given the studies are longitudinal, in the cohorts that provide feedback of abnormal results at biomedical sweeps, undiagnosed cases may subsequently become diagnosed.

Less is known about the numbers who do not report a diagnosed condition (i.e., false negatives), and so diagnosed prevalence using self-reports only may be an underestimate.

2.3.2 Changes over time

Changes in diagnosed cases by time and cohort can be driven by both changes in diabetes prevalence, and by changes in diagnostic tests and screening resulting in more cases receiving diagnoses. The gold standard for diabetes diagnosis remains testing plasma glucose concentration and oral glucose tolerance tests. However, in more recent years, a glycated haemoglobin (HbA1c) concentration ≥48 mmol/mol can also be used to diagnose diabetes (Hitman, 2012). While an HbA1c reading below the threshold does not rule out diabetes, the comparative ease of testing HbA1c concentration (finger-prick test) has likely contributed to a larger number of people being diagnosed with diabetes. The introduction of diabetes screening programmes, such as the National Health Service Health Check (introduced in 2009), and of clinical targets in primary care such as the Quality and Outcomes Framework, have also likely resulted in an increase in the proportion of diabetes cases that are diagnosed (especially for type 2 diabetes). Thus, there may be more undiagnosed cases in the older cohorts in earlier years.

2.3.3 Type of diabetes

It is challenging to identify the type of diabetes consistently across studies. There are few questions in the cohorts which specifically ask about type of diabetes. The terminology used to describe diabetes type has changed over time and across cohorts (e.g., sugar diabetes, insulin-dependent diabetes, type 1 diabetes, type 2 diabetes). Therefore, classifying diabetes type within studies requires assumptions, which have been documented in this report. Cohort members may be unsure of the type of diabetes that they have. This is reflected both in responses to direct questions about diabetes type, where many respond that they do not know their diabetes type, and in responses to longstanding illness questions which when coded up to ICD-10 were most often non-specific diabetes (E12, E13, or E14) rather than E10 (type 1 diabetes) or E11 (type 2 diabetes). Additionally, we found cases where the type of diabetes self-reported by the respondent did not seem to match up with other information provided (e.g., insulin use, age at diagnosis). The proportion of cases of self-reported type 1 diabetes cases that were flagged for caution was

declined across cohorts (being lowest in the 1970 cohort), potentially reflecting growing awareness of diabetes across cohorts.

Generally, we can be confident that diabetes diagnosed in childhood and adolescence is type 1 diabetes and diabetes diagnosed in midlife and older is type 2 diabetes. However, disentangling type outside of these age-ranges with limited self-reported data on type is challenging, especially given differences the potentially declining age of type 2 diabetes onset in more recently born cohorts (Holden *et al.,* 2013), and a likely higher prevalence of late type 1 diabetes onset or diagnosis in older cohorts (Dahlquist *et al.,* 2011). Misclassification of type among those who were assigned based on information other than self-reported type is therefore also possible.

Finally, in cohorts where there is very limited data on diabetes (Next Steps and MCS), it was not possible to assign a diabetes type. Linkage to electronic health records will enable validation of our classification of diabetes type.

3. Description of the Harmonised Datasets

3.1 Licensing and data access

The harmonised data for diabetes has been processed by CLS and supplied to the UK Data Service. All data users need to be registered with the UK Data Service and to sign the UKDS End User Licence before they can download the data. Details of how to do this are available at ukdataservice.ac.uk/get-data/how-to-access/registration.

The NCDS, BCS70, Next Steps and MCS datasets are available as safeguarded data, which can be downloaded from the UK Data Service once the End User Licence (EUL) access conditions have been accepted by the user.

The NSHD dataset can be accessed by downloading the UKDS Special Licence application form. Once the form has been reviewed by UKDS and approved by the NSHD Data Sharing Committee the data will be available to download. For accessing and linking to other NSHD data see section 3.2.

Access for additional NSHD data

The NSHD diabetes dataset is also available from MRC Unit for Lifelong Health and Ageing at UCL (LHA), which manages the NSHD. This route of access is necessary for analysts wishing to use the diabetes data alongside other information held for the 1946 cohort. The research project needs to first be approved by the NSHD Data Sharing Committee. Full details on how to access the data can be found here. Once a data access form has been approved and a data sharing agreement is in place, the data can be accessed via www.condor.ucl.ac.uk .

3.2 List of datasets

Table 1: List of available harmonised datasets

Name of the dataset	Content summary
harmonised_diabetes_nshd	Harmonised indicators of self-reported diabetes in the 1946 NSHD. Includes sweep-specific indicators, derived cumulative indicators, and derived diabetes type.
harmonised_diabetes_ncds	Harmonised indicators of self-reported diabetes in the 1958 NCDS. Includes sweep-specific indicators, derived cumulative indicators, and derived diabetes type.
harmonised_diabetes_bcs	Harmonised indicators of self-reported diabetes in the 1970 BCS. Includes sweep-specific indicators, derived cumulative indicators, and derived diabetes type.
harmonised_diabetes_nextsteps	Harmonised indicators of self-reported diabetes in Next Steps. Includes sweep-specific indicators only.
harmonised_diabetes_mcs	Harmonised indicators of self-reported diabetes in MCS. Includes sweep-specific indicators only.

All datasets have a flat/wide structure, with one row per cohort member.

3.3 Harmonised variables syntax

The code developed to derive all harmonised diabetes variables can be found on the CLS Data GitHub page https://github.com/CLS-Data/harmonised-diabetes-across-cohorts

3.4 Identifiers

Individual identifiers

Table 2: Individual identifiers

Name of the dataset	Individual identifier
harmonised_diabetes_nshd	nshdid_ukds01
harmonised_diabetes_ncds	ncdsid
harmonised_diabetes_bcs	bcsid
harmonised_diabetes_nextsteps	nsid
harmonised_diabetes_mcs	mcsid (MCS family identifier) and cnum (individual person identifiers).

Use of individual identifiers to merge with cohort study data

For NCDS, BCS70, Next Steps and MCS, the data are identified with the same research IDs used for the rest of cohort data available at the UK Data Service. This enables the data to be merged with one another.

For MCS, researchers need to use both the MCS family identifier (MCSID) and the individual person identifier (CNUM) to merge on with other cohort data.

For NSHD, research IDs differ from the original research IDs and data can therefore not be merged with other NSHD data files. Access to the harmonised variables in connection with other NSHD variables can be requested via usual the NSDH data sharing request as all harmonised asthma variables are also included in the NSDH data catalogue.

3.5 Variable Naming Conventions

Within each dataset, variable naming conventions have been used to indicate to users the primary source of information on diabetes.

Variable naming conventions have been used to indicate to users the primary source of information on diabetes.

Sweep-specific indicators

All indicators of presence of diabetes at a given sweep are named **diab_xx_y**, where:

- 'xx' indicates the cohort member's age at the sweep.
- 'y' indicates the primary source of information on diabetes:
    - 'ever' indicates that the information comes from answers to an 'ever diabetes' question, which may further be enriched through responses to longstanding illness type questions at the same sweep.
    - 'current' indicates that the information primarily comes from answers to a 'current diabetes' question, which may further be enriched through responses to longstanding illness type questions at the same sweep.
    - 'doc' is used to indicate a small number of cases in childhood sweeps where information comes from doctors' assessments of cohort members' medical history (instead of parental reports). They can be interpreted similarly to the 'current' indicators, though note that there is often more missing data among those who participated in the sweep, since not all cohort members underwent medical examination.

For each of these indicators, the following response categories are possible:

- 1 = Diabetes reported in sweep
- 2 = No diabetes reported in sweep
- -1 = No information provided
- -9 = Not in sweep.

Cross-sweep indicators

Cumulative indicators were only derived for NSHD, NCDS and BCS70, since the only information on diabetes currently available for MCS and Next Steps comes from the COVID-19 sweeps.

- Cumulative ‘ever’ indicators are named **diab_xx_cumul** where xx refers to the age at the sweep up to which the indicator is cumulated. For each indicator, the following response categories are possible:
    - 1 = Diabetes ever reported up to and including current sweep
    - 2 = No diabetes reported up to and including current sweep
    - -1 = Not enough information provided
    - -9 = Not in sweep.

- Derived diabetes type is recorded in a variable named **diab_type.** For each indicator, the following response categories are possible:
    - 1 = Type 1
    - 2 = Type 2
    - 3 = Gestational diabetes only
    - 4 = Undetermined type
    - -1 = Diabetes status unknown
    - -3 = Not applicable (never reported diabetes).

- The variable **diab_type_rsn** contains additional information about why an individual was classified as a particular type. For each indicator, the following response categories are possible:
    - 1 = From self-reported type
    - 2 = Childbirth within 1 year of age at diagnosis
    - 3 = From reported use of insulin
    - 4 = From age at diagnosis/insulin <20 or >40
    - 5 = From first report in sweep aged <20 or >40
    - 6 = From patterning of responses to current diabetes questions
    - -1 = Diabetes status unknown
    - -3 = Not applicable (never reported diabetes)
    - -4 = Not enough information to assign type.

- The variable **diab_type_caution** indicates whether cohort members report insulin use or age at diagnosis which raises questions over the diabetes type they have self-reported. It can take the following values:

    - 1 = Caution flag raised
    - 2 = No caution flag raised
    - -1 = Diabetes status unknown
    - -3 = Not applicable (never reported diabetes)
    - -4 = Not applicable (no type assigned).

- The variable **diab_type_caution_rsn** holds information on why a cohort member has a caution flag for diabetes type. It can take the following values:
    - 1 = Self-reported type 1 but reports no insulin use
    - 2 = Self-reported type 1 but age at diagnosis > 30 years
    - 3 = Self-reported type 2 but age at diagnosis < 20 years
    - -1 = Diabetes status unknown
    - -2 = Not applicable (no caution flag)
    - -3 = Not applicable (never reported diabetes)
    - -4 = Not applicable (no type assigned).

3.6 Accessing Intermediate Variables

The deposited dataset only retains the derived harmonised indicators for diabetes, however, throughout the process of deriving these indicators we have documented or derived other available self-reported information on diabetes across the cohorts (such as variables based on longstanding illness type questions). Researchers who are interested can derive these indicators themselves using the code we have deposited (https://github.com/CLS-Data/harmonised-diabetes-across-cohorts). Researchers interested in using the 1946 cohort will need to request relevant variables from them to run the code.

4. References

Bann, D., Wright, L., Goisis, A., Hardy, R., Johnson, W., Maddock, J., McElroy, E., Moulton, V., Patalay, P., Scholes, S., Silverwood, R.J., Ploubidis, G.B., O'Neill, D. (2022) Investigating change across time in prevalence or association: The challenges of cross-study comparative research and possible solutions. *Discover Social Science and Health,* 2(18). https://doi.org/10.1007/s44155-022-00021-1

Chen, S., Zhang, Q., Dai, G., Hu, J., Zhu, C., Su, L., Wu, X. (2016) Association of diabetes with pre-diabetes, undiagnosed diabetes, and previously diagnosed diabetes: A meta-analysis. *Endocrine, 53*: 35-46. https://doi.org/10.1007/s12020-016-0869-x

Connelly, R. & Platt, L. (2014). Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology,* 43(6): 1719-1725. https://doi.org/10.1093/ije/dyu001

Dahlquist, G.G., Nyström, L., Patterson, C.C., the Swedish Childhood Diabetes Study Group & the Diabetes Incidence in Sweden Study Group (2011). Incidence of Type 1 Diabetes in Sweden Among Individuals Aged 0-34 Years, 1983-2007. *Diabetes Care,* 34(8): 1754-1759.

Diabetes UK. (n.d.) How many people in the UK have diabetes? [Accessed 9 March 2024]. https://www.diabetes.org.uk/about-us/about-the-charity/our-strategy/statistics

Elliot, J. & Shepherd, P. (2006) Cohort Profile: 1970 British Cohort Study (BCS70). *International Journal of Epidemiology,* 35(4): 836-843. https://doi.org/10.1093/ije/dyl174

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L.E., Craig, C., Saliba, M., … Ferretti, V. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonisation. *International Journal of Epidemiology, 46*(1): 103-105. https://doi.org/10.1093/ije/dyw075

Gondek, D. (2020) *We are living longer, but not healthier: Evidence from the British birth cohorts and the Uppsala Birth Cohort Multigenerational Study*. Doctoral Thesis. London: University College London.

Hackett, R. A., Moore, C., Steptoe, A., Lassale, C. (2018) Health behaviour changes after type 2 diabetes diagnosis: Findings from the English Longitudinal Study of Ageing. *Scientific Reports, 8:* 16938. https://doi.org/10.1038/s41598-018-35238-1

Hitman, G.A. (2012) Finally, a UK consensus on the use of HbA1c to diagnose diabetes. *Diabetic Medicine,* 29: 1349. https://doi.org/10.1111/dme.12011

Holden, S.E., Barnett, A.H., Peters, J.R., Jenkins-Jones, S., Poole, C.D., Morgan, C.L., Currie, C.J. (2013) The incidence of type 2 diabetes in the United Kingdom from 1991 to 2010. *Diabetes, Obesity and Metabolism;* 15(9): 844-852. https://doi.org/10.1111/dom.12133

Pastorino, S., Richards, M., Hardy, R., Abington, J., Wills, A., Kuh, D., Pierce, M., and the National Survey of Health and Development Scientific and Data Collection Teams. (2015) Validation of self-reported diagnosis of diabetes in the 1946 British birth cohort. *Primary Care Diabetes, 9*(5): 397-400. https://doi.org/10.1016/j.pcd.2014.05.003

Power, C. & Elliott, J. (2006) Cohort profile: The 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology,* 35(1): 34-41. https://doi.org/10.1093/ije/dyi183

Sullivan, A., Brown, M., Hamer, M., Ploubidis, G.B. (2023) Cohort Profile Update: The 1970 British Cohort Study (BCS70). *International Journal of Epidemiology,* 52(3): e179-e186. https://doi.org/10.1093/ije/dyac148

Vounzoulaki, E., Khunti, K., Tan, B.T., Gillies, C.L. (2020) Progression to type 2 diabetes in women with a known history of gestational diabetes: systematic review and meta-analysis. *British Medical Journal;* 369:m1361. https://doi.org/10.1136/nmj.m1361

Wadsworth, M., Kuh, D., Richards, M., Hardy, R. (2006) Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology,* 35(1): 49-54.

Appendix 1: Datasets Used for Harmonisation

1946 National Study of Health and Development

All variables for NSHD were requested through Skylark. The variables included in the derivation can be seen in Appendix 2. Users should also request response variables if they plan to derive the harmonised outcomes.

1958 National Child Development Study

All datasets used are available via the UK Data Service to registered users, through an end user license agreement:

- Response file: SN 5560
- Age 0-16: SN 5565
- Age 23: SN 5566
- Age 33: SN 5567
- Age 42: SN 5578
- Age 46: SN 5579
- Age 50: SN 6137
- Age 55: SN 7669
- COVID-19 sweeps: SN 8658.

1970 British Cohort Study

All datasets used are available via the UK Data Service to registered users, through an end user license agreement:

- Response file: SN 5641
- Age 10: SN 3723
- Age 26: SN 3833
- Age 30: SN 5558
- Age 34: SN 5585
- Age 38: SN 6557
- Age 42: SN 7473
- Age 46: SN 8547
- Age 51: SN 9347

- COVID-19 sweeps: SN 8658.

Next Steps

All datasets used are available via the UK Data Service to registered users, through an end user license agreement:

- All Next Steps sweeps: SN 5545
- COVID-19 sweeps: SN 8658.

Millennium Cohort Study

All datasets used are available via the UK Data Service to registered users, through an end user license agreement:

- Response file: SN 8172
- COVID-19 sweeps: SN 8658.

Appendix 2: Indicators and Variables Used to Derive Harmonised Measures

1946 MRC National Survey of Health and Development

If you would like to derive these variables yourself, then please note that it is also important to request the following variables from NSHD, to ensure that missingness is coded appropriately across sweeps.

Age	Year	Question	Original name	New name	Coding of new variable
Sweep-specific indicators					
36	1982	Do you have any of the following most or all of the time? Diabetes. Self-report.	diab82	diab_36_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
43	1989	Have you ever had any of the following? Diabetes. Self-report.	diab89	diab_43_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
53	1999	In the last 10 years (since you were 43 years old) have you had diabetes? Self-report.	diab	diab_53_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
63	2009	Since 1999 have you been told that you have diabetes? Self-report. Has a doctor told you that you have any of the following health problems? Self-report. Since 199 have you suffered from any other troublesome health problems which have been diagnosed by a medical doctor? If Yes, please list below. Self-report.	diab09 ddiab09 othpb_60, othhp109, othhp209, othhp309, othhp409, othhp509, othhp609, othhp709, othhp809, othhp909, othhp1009, othhp1109	diab_63_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Age	Year	Question	Original name	New name	Coding of new variable
68	*2014*	*Since 2006 have you been told by a doctor that you have diabetes? [Postal]. Self-report.*	*diab14x*	*diab_68_ever*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
69	*2015*	*Do you have diabetes? [Nurse visit]. Self-report.* *Do you have any other diagnosed medical condition not already mentioned? What is it? What was your age when it was diagnosed? [Nurse visit]. Self-report.*	*ddiab15x* *ddiab215x* *othhp115x, othhp215x, othhp315x, othhp415x, othhp515x, othhp615x, othhp715x*	*diab_69_current*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
74	*2020*	*Do you have any of the following? [COVID Sweep 1]. Self-report.*	*cw1_lli_6*	*diab_74_current1*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
74	*2020*	*Do you have any of the following? Diabetes. [COVID sweep 1] Self-report.* *Do you have any of the following? Diabetes. [COVID Sweep 2]. Self-report.*	*cw1_lli_6* *cw2_lii1_6*	*diab_74_current2*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
75	*2021*	*Do you currently have any of the following? Diabetes. [COVID Sweep 3]. Self-report.*	*cw3_lli1_6*	*diab_75_cur rent*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
76	*2022*	*Since your 68th birthday have you been told that you have diabetes? Self-report.*	*diab22x*	*diab_76_ever*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Age	Year	Question	Original name	New name	Coding of new variable
Cross-sweep indicators					
36	-	Ever reported diabetes by age 36 [cumulative]	-	diab_36_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
43	-	Ever reported diabetes by age 43 [cumulative].	-	diab_43_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
53	-	Ever reported diabetes by age 53 [cumulative].	-	diab_53_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
63	-	Ever reported diabetes by age 60-64 [cumulative].	-	diab_63_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
68	-	Ever reported diabetes by age 68 [cumulative].	-	diab_68_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep

Age	Year	Question	Original name	New name	Coding of new variable
					-1 = Not enough information provided -9 = Not in sweep
69	-	*Ever reported diabetes by age 69 [cumulative].*	-	*diab_69_cumul*	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
75	-	*Ever reported diabetes by age 75 [cumulative, end of COVID sweeps].*	-	*diab_75_cumul*	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
76	-	*Ever reported diabetes by age 76 [cumulative].*	-	*diab_76_cumul*	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
-	-	*Type of diabetes*	-	*diab_type*	1 = Type 1 2 = Type 2 3 = Gestational diabetes only 4 = Undetermined type -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes)
-	-	*Reason diabetes type was assigned*	-	*diab_type_rsn*	1 = From self-reported type

Age	Year	Question	Original name	New name	Coding of new variable
					2 = Childbirth within 1 year of age at diagnosis 3 = From reported use of insulin 4 = From age at diagnosis or insulin use <20 or >40 5 = From first report in sweep <20 or >40 6 = From patterning of responses to current diabetes questions -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes) -3 = Not applicable (no type assigned)
-	-	*Caution with self-reported diabetes type*		*diab_type_caution*	1 = Caution flag raised 2 = No caution flag raised -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes) -3 = Not applicable (no type assigned)
-	-	*Reason for caution with self-reported diabetes type*	-	*diab_type_caution_rsn*	1 = Self-reported type 1 but reports no insulin use 2 = Self-reported type 1 but age at diagnosis > 30 years 4 = Self-reported type 1 but no insulin and age at diagnosis > 30 years 4 = Self-reported type 2 but age at diagnosis < 20 years -1 = Diabetes status unknown -2 = Not applicable (no caution flag)

Age	*Year*	*Question*	*Original name*	*New name*	*Coding of new variable*
					-3 = Not applicable (never reported diabetes) -4 = Not applicable (no type assigned)

In the derivation of diabetes type, we also used information on childbirth (chay182_v2, chay282_v2, chay382_v2 and chay482_v2), and previously derived information on insulin use (medinsulin_31x, medinsulin_36x, medinsulin_43x, medinsulin_53x, medinsulin_63x, and medinsulin_69x). We used variables recording sweep response to code missing values (int36rec, intr43rec, int53rec, int63rec, int68rec, int69rec, int_cw1rec, int_cw2rec, int_cw3rec, int76rec). All variables can be accessed through NSHD's Skylark platform.

1958 National Child Development Study

Age	Year	Question	Original name	New name	Coding of new variable
7	1965	Summary of abnormal conditions at medical exam (diabetes). *Doctor report.*	n417	diab_7_currentdoc	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
16	1974	Having completed the examination [does the cohort member have … and to what degree does it create a handicap]?" (diabetes). *Doctor report.* Taking into account the information you have obtained during the interview and other relevant information, do you consider the child has any handicapping conditions or disability? If yes, what is the nature of the child's handicap or disability? *Interviewer report.*	n2034 n2662, n2663	diab_16_currentdoc	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep **Note that item nonresponse is high since not all cohort members in the sweep participated in the medical examination.**
33	1991	Have you ever suffered from or been told you had diabetes? *Self-report.* Including any health problems you may have already told me about do you have any longstanding illness, disability, or infirmity of any kind? What is the name of this condition? *Self-report.*	n503921 n509031, n509034, n509037, n509040	diab_33_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
33	1991	Have you suffered or been told you suffer from diabetes in the last 12 months? *Self-report.*	n503922 n509031, n509034, n509037, n509040	diab_33_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
42	2000	Have you ever had or been told you had diabetes? *Self-report.*	diab	diab_42_ever	1 = Diabetes reported in sweep

Age	Year	Question	Original name	New name	Coding of new variable
					2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
42	2000	Have you had diabetes in the last 12 months? *Self-report.*	dl112m	diab_42_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
50	2008	Are you currently suffering from any other health problems listed on this card? (Sugar) diabetes. *Self-report (or proxy report)* Apart from those you have already told us about, do you have any [other] longstanding illness, disability, or infirmity? What is the matter with you? *Self-report (or proxy report)*	N8KHPB03, N8XKHP03 N8XLSA01-N8XLSA11, N8XXLSA1-N8XXLSA3	diab_50_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
55	2013	Since [date of last interview/5 years ago] have you had any of the following conditions? (Diabetes). *Self-report.* We inform this variable with previous responses about diabetes, building from the last "ever" question at age 42: If diab_42_e = 1 or diab_50_c = 1, then diab_55_e = 1.	N9KHLPRB02	diab_55_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
62	2020	Do you have any of the following? (Diabetes). *Self-report.* [COVID Sweep 1]	CW1_LLI6	diab_62_current1	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
62	2020	Do you have any of the following? (Diabetes). *Self-report.* [COVID Sweep 1] Do you have any of the following? (Diabetes). *Self-report.* [COVID Sweep 2]	CW1_LLI6 CW2_LLI1_6	diab_62_current2	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided

Age	Year	Question	Original name	New name	Coding of new variable
					-9 = Not in sweep
63	2021	Do you have any of the following? (Diabetes). *Self-report.* [COVID Sweep 3]	CW3_LLI_6	diab_63_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
Cross-sweep indicators					
16	*1974*	*Ever reported diabetes by age 16 [cumulated, sweep 3].*	-	diab_16_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
33	*1991*	*Ever reported diabetes by age 33 [cumulated, sweep 5].*	-	diab_33_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
42	*2000*	*Ever reported diabetes by age 43 [cumulated, sweep 6].*	-	diab_42_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
50	*2008*	*Ever reported diabetes by age 50 [cumulated, sweep 8].*	-	diab_50_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep

Age	Year	Question	Original name	New name	Coding of new variable
					-1 = Not enough information provided -9 = Not in sweep
55	*2013*	*Ever reported diabetes by age 55 [cumulated, sweep 9].*	-	diab_55_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
63	*2021*	*Ever reported diabetes by age 63 [cumulated, end of COVID sweeps].*	-	diab_63_cumul	1 = Yes 0 = No .m = status unknown (don't know, refused, didn't answer, etc) . = unit nonresponse (not in sweep)
-	-	*Type of diabetes*	-	*diab_type*	1 = Type 1 2 = Type 2 3 = Gestational diabetes only 4 = Undetermined type -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes)
-	-	*Reason diabetes type was assigned*	-	*diab_type_rsn*	1 = From self-reported type 2 = Childbirth within 1 year of age at diagnosis 3 = From reported use of insulin 4 = From age at diagnosis or insulin use <20 or >40 5 = From first report in sweep <20 or >40

Age	Year	Question	Original name	New name	Coding of new variable
					6 = From patterning of responses to diabetes questions -1 = Diabetes status unknown -3 = Not applicable (never reported diabetes) -4 = Not applicable (no type assigned)
-	-	*Caution with self-reported diabetes type*		*diab_type_caution*	1 = Caution flag raised 2 = No caution flag raised -1 = Diabetes status unknown -3 = Not applicable (never reported diabetes) -4 = Not applicable (no type assigned)
-	-	*Reason for caution with self-reported diabetes type*	-	*diab_type_caution_rsn*	1 = Self-reported type 1 but 1 = Self-reported type 1 but reports no insulin use 2 = Self-reported type 1 but age at diagnosis > 30 years 4 = Self-reported type 1 but no insulin and age at diagnosis > 30 years 4 = Self-reported type 2 but age at diagnosis < 20 years -1 = Diabetes status unknown -2 = Not applicable (no caution flag) -3 = Not applicable (never reported diabetes) -4 = Not applicable (no type assigned)

1970 British Cohort Study

Age	Year	Question	Original name	New name	Coding of new variable
26	1996	Since you were 16, have you suffered from diabetes? *Self-report.* Do you suffer from any long-term health problems, longstanding illness, infirmity, or disability of any kind? If yes, please describe. *Self-report.*	b960455 b960566, q41oth*	diab_26_ever	1 = Diabetes since age 16 2 = No diabetes since age 16 -1 = No information provided -9 = Not in sweep
26	1996	If you have suffered from diabetes since you were 16, was this in the last 12 months? *Self-report.* Do you suffer from any long-term health problems, longstanding illness, infirmity, or disability of any kind? If yes, please describe. *Self-report.*	b960521 b960566, q41oth*	diab_26_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
30	2000	Have you ever had or been told you had diabetes? *Self-report.*	diab	diab_30_ever	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
30	2000	Have you had diabetes in the last 12 months? *Self-report.*	dl112m	diab_30_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
34	2004	Since [last interview/January 2000] have you had any of the health problems listed on this card? (Sugar) diabetes. *Self-report.* Do you have any longstanding illness, disability, or infirmity? What [else] is the matter with you? *Self-report.*	b7hpb04 b7xlsa-b7xlsa7, b7xlsb-b7xlsb6, b7xlsc-b7xlsc3,	diab_34_ever **The recoded version of this variable carries forward information from age 30.**	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Age	Year	Question	Original name	New name	Coding of new variable
		Any other health problems or difficulties? *Self-report.*	b7xlsd-b7xlsd2, b7xlse b7xkhlba, b7xkhlbb, b7xkhlbc, b7xkhlbd		
38	2008	Can you tell me whether you are currently suffering from any of the following conditions? (Sugar) diabetes. *Self-report.*	b8khpb03	diab_38_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
42	2012	Since [last interview/2008] have you had any of the health problems on this card? (Sugar) diabetes. *Self-report.*	B9KHPB04, B9PKHP04	diab_42_ever **The recoded version of this variable carried forward information on diabetes at ages 30, 34 and 38.**	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
46	2016	[Since last interview/in the last four years] have you had any of the health problems listed on this card? Please include any problems that had already started before that date. (Diabetes). *Self-report.*	B10KHPB03	diab_46_ever **The recoded version of this variable carried forward information on diabetes at ages 30, 34, 38 and 42.**	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
50	2020	Do you have any of the following? (Diabetes). *Self-report COVID sweep.*	CW1_LLI6	diab_50_current1	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Age	Year	Question	Original name	New name	Coding of new variable
50	2020	Do you have any of the following? (Diabetes). *Self-report COVID sweep*.	CW1_LLI6 CW2_LLI1_6	diab_50_current2	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
51	2021	Do you have any of the following? (Diabetes). *Self-report COVID sweep*.	CW3_LLI_6	diab_51_current	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
51-54	2021-2024	[Since last interview/in the last four years] have you had any of the health problems listed on this card? (Sugar) diabetes. *Self-report*.	b11khpb04	diab_54_ever **The recoded version of this variable carried forward information on diabetes at ages 30, 34, 38 and 42.**	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
Cross-sweep indicators					
26	1996	*Ever reported diabetes by age 26 [cumulated, sweep 5].*	-	diab_26_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
30	2000	*Ever reported diabetes by age 30 [cumulated, sweep 6].*	-	diab_30_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided

Age	Year	Question	Original name	New name	Coding of new variable
					-9 = Not in sweep
34	2004	*Ever reported diabetes by age 34 [cumulated, sweep 7]*	-	diab_34_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
38	2008	*Ever reported diabetes by age 38 [cumulated, sweep 8]*	-	diab_38_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
42	2012	*Ever reported diabetes by age 42 [cumulated, sweep 9]*	-	diab_42_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
46	*2016*	*Ever reported diabetes by age 46 [cumulated, sweep 10]*	-	diab_46_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
51	*2021*	*Ever reported diabetes by age 51 [cumulated, end of COVID sweeps]*	-	diab_51_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep

Age	Year	Question	Original name	New name	Coding of new variable
					-1 = Not enough information provided -9 = Not in sweep
51	*2021*	*Ever reported diabetes by age 51-54 [cumulated, sweep 11]*		diab_54_cumul	1 = Diabetes ever reported up to and including current sweep 2 = No diabetes ever reported up to and including current sweep -1 = Not enough information provided -9 = Not in sweep
-	-	*Type of diabetes*	-	*diab_type*	1 = Type 1 2 = Type 2 3 = Gestational diabetes only 4 = Undetermined type -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes)
-	-	*Reason diabetes type was assigned*	-	*diab_type_rsn*	1 = From self-reported type 2 = Childbirth within 1 year of age at diagnosis 3 = From reported use of insulin 4 = From age at diagnosis or insulin use <20 or >40 5 = From first report in sweep <20 or >40 6 = From patterning of responses to current diabetes questions -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes) -3 = Not applicable (no type assigned)

Age	Year	Question	Original name	New name	Coding of new variable
-	-	*Caution with self-reported diabetes type*		*diab_type_caution*	1 = Caution flag raised 2 = No caution flag raised -1 = Diabetes status unknown -2 = Not applicable (never reported diabetes) -3 = Not applicable (no type assigned)
-	-	*Reason for caution with self-reported diabetes type.*	-	*diab_type_caution_rsn*	1 = Self-reported type 1 but reports no insulin use 2 = Self-reported type 1 but age at diagnosis > 30 years 4 = Self-reported type 1 but no insulin and age at diagnosis > 30 years 4 = Self-reported type 2 but age at diagnosis < 20 years -1 = Diabetes status unknown -2 = Not applicable (no caution flag) -3 = Not applicable (never reported diabetes) -4 = Not applicable (no type assigned)

Next Steps

Age	Year	Question	Original name	New name	Coding of new variable
Sweep-specific indicators					
30	*2020*	*Do you have any of the following? Diabetes. [COVID Sweep 1]. Self-report.*	*CW1_LLI_6*	*diab_30_current1*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
30	*2020*	*Do you have any of the following? Diabetes. [COVID Sweep 1]. Self-report.* *Do you have any of the following? Diabetes. [COVID Sweep 2]. Self-report.*	*CW1_LLI_6* *CW2_LLI1_6*	*diab_30_current2*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
31	*2021*	*Do you have any of the following? Diabetes. [COVID Sweep 3]. Self-report.*	*CW3_LLI_6*	*diab_31_current*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Millennium Cohort Study

Age	Year	Question	Original name	New name	Coding of new variable
Sweep-specific indicators					
20	*2020*	*Do you have any of the following? Diabetes. [COVID Sweep 1]. Self-report.*	*CW1_LLI_6* *CW2_LLI1_6*	*diab_20_current1*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
20	*2020*	*Do you have any of the following? Diabetes. [COVID Sweep 1]. Self-report.* *Do you have any of the following? Diabetes. [COVID Sweep 2]. Self-report.*	*CW1_LLI_6*	*diab_20_curent2*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep
21	*2021*	*Do you have any of the following? Diabetes. [COVID Sweep 3]. Self-report.*	*CW3_LLI_6*	*diab_21_current*	1 = Diabetes reported in sweep 2 = No diabetes reported in sweep -1 = No information provided -9 = Not in sweep

Appendix 3: Agreement of harmonised NHSD variables

Another set of variables related to diabetes is available for the 1946 NSHD. These variables were derived based on medical records and survey responses up to the age 53 sweep (**diabre** and **diabty**).

The harmonised variables provided here offer an updated version that includes information from sweeps up to age 76. Owing to coding decisions which aim to make the variables produced for NSHD comparable with those produced for other cohorts, there are some minor discrepancies between the harmonised indicators and **diabec** and **diabty**. We have briefly quantified these disagreements below.

Table 3.1. Comparison of cumulated diabetes prevalence up to and including age 53 sweep from harmonisation project and ever diabetes indicator *diabec*.

	Harmonisation project	***diabec***
Overall	2.94%	2.72%
Male	3.07%	3.08%
Female	2.83%	2.36%

Note: Estimates from the harmonised data are based on those who responded to the age 53 sweep only, and individuals could be "back-coded" to having diabetes at age 53 if they subsequently reported an age of diabetes onset ≤ 53 years. At age 53, n = 3023.

Disagreements between *diabec* and the harmonised cumulated indicator at age 53 were low (<5 cases of recorded in *diabec* at age 53 were missed using the harmonised indicator).

Table 3.2. Diabetes type as assigned using harmonisation algorithm compared to diabetes type indicator *diabty*.

	Harmonisation Project					
diabty	**None**	**Type 1**	**Type 2**	**Gestational**	**Unknown**	**Total**
Type 1		70%	15%		15%	<15 (100%)
Type 2		20%	80%			68 (100%)
Gestational	30%		70%			<5 (100%)
Unknown	50%	25%			25%	<5 (100%)
Unknown likely Type 1		90%	10%			<10 (100%)
Unknown likely Type 2			50%		50%	<5 (100%)

Note: Discrepancies between *diabty* and our Type 1/Type 2 assignments are driven by cohort member's self-reported diabetes type, which takes first order of precedence in our algorithm.