DD2

SKNC

Invalid Date

Table of contents

Chapter 1

The Danish Center for Strategic Research in Type 2 Diabetes

❗ Important

This website is currently under development

This website includes details about **The Danish Center for Strategic Research in Type 2 Diabetes** (The **DD2** Study). The intention of this website is to document the data sources used for research and to support the day-to-day work with DD2. It is mainly aimed at employees of the Department of Clinical Epidemiology (DCE/KEA), Aarhus University and Aarhus University Hospital, as it describes the setup used at DCE. Other institutions can also access DD2 data but their setup might differ.

To learn more about DD2 please see: https://dd2.dk/

Please see the 2012 paper about the DD2 enrollment process (Nielsen et al. (2012)) and the 2018 and 2024 cohort papers (Christensen et al. (2018) and Kristensen et al. (2024)) with information about enrollment, data sources, and baseline characteristics. An extensive list of scientific publications based on data from DD2 is available on the official website.

Chapter 2

DD2 enrollment

Enrollment in DD2 includes a questionnaire and the collection of urine and blood samples. The questionnaire includes a section with a few questions for the patient to fill out (at registration for enrollment) and another to be filled out by the enroller (at enrollment). All individuals older than 18 years of age diagnosed with diabetes within the last 2 years are eligible for inclusion in DD2 (previously, a longer diabetes duration was allowed).

Individuals may either sign up online and be contacted later, or be approached by practice personnel during routine care. The individual will receive an e-mail with a link to WebPatient, where a few questions should be filled out before the questionnaire will be sent to the practice. Hereafter, a DD2 examination will take place, where additional questions will be answered (e.g., a clinically assessed date of diabetes onset) along with a physical examination (e.g., to measure weight, hip and waist circumference, and resting heart rate). The examination also involves a urine sample (morning urine) and a blood sample (fasting).

Enrollment can either take place in the general practice (GP) or in the outpatient clinic. Registration is also possible for individuals without internet access. For further information about the enrollment, please see the DD2 website (Danish): Patient information and Instructions and questionnaires.

The first individual was enrolled in DD2 in November 2010. Between 2010 and April 2023, the DD2 cohort enrolled 11,381 individuals. An additional 717 individuals were enrolled by November 2023, yielding a total of N=12,098 individuals. In the 2024 cohort paper (Kristensen et al. (2024)), we illustrated the enrollment during the period from November 2010 to April 2023.

The median age at enrollment was 61 years (IQR 53–69 years) and 41% were women (Kristensen et al. (2024)). Individuals are eligible for inclusion in

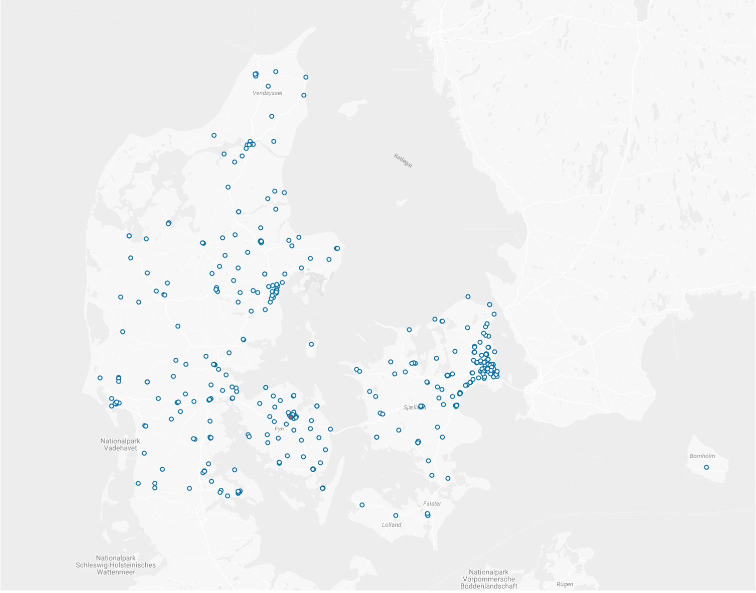
Vendsyssel
Kattegat
Nationalpark Vadehavet
Fyn
Sjælland
Bornholm
Falster
Lolland
Nationalpark Schleswig-Holsteinisches Wattenmeer
Nationalpark Vorpommersche Boddenlandschaft
Rügen

Figure 2.1: Figure from https://dd2.dk/om-dd2/hvem-er-med (downloaded 30 October 2025) showing the general practitioners and outpatient clinic enrolling individuals in DD2.

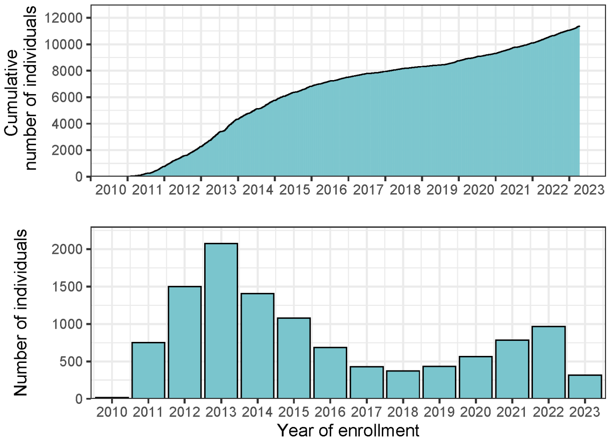

Figure 2.2: Figure from Kristensen et al. (2024) with a total of N=11,369 individuals. Overview of enrollment in the DD2 cohort from November 2010 to April 2023. The plots show the cumulative and annual number of enrolled individuals.

DD2 if they are diagnosed with diabetes within the last 2 years (a longer diabetes duration was allowed in the early phase). Median diabetes duration at time of enrollment was 1.3 years (IQR 0.4–2.9 years) and 87% were already prescribed with glucose-lowering treatment at the time of enrollment (Kristensen et al. (2024)).

For more information about the enrollment and data collection, please contact DD2. Some of the general DD2 documents can also be downloaded here (Danish, downloaded 12 January 2024):

- DD2 questionnaire:
- Inclusion information:
- Biobank samples: and

Chapter 3

Data structure

The full DD2 research data are based on data collected and linked from multiple sources (Kristensen et al. (2024)). This page gives an overview of the data structure, the data sources, and the definitions of data types. Detailed descriptions of each data source are provided on the following pages.

Within the DD2 data collection, the data sources include **primary data**, i.e., data unique to DD2, including data and material collected from participants at enrollment, and additional data from research studies within the DD2 cohort. The data also include **secondary data**, i.e., external data collected from other data sources and then linked to DD2. Data from the DD2 data collection are also linked to the nationwide routine registry data located in centralized databases on remote research servers (SDS and DST).

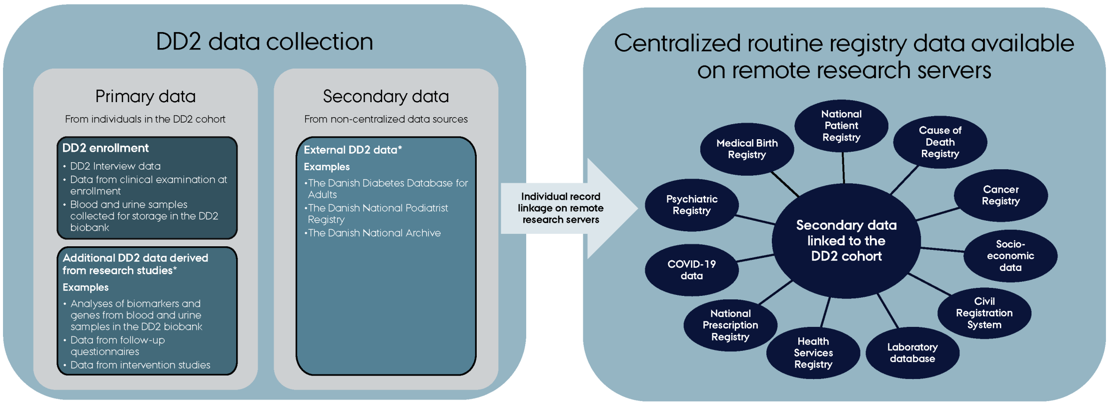
DD2 data collection
Primary data
From individuals in the DD2 cohort
DD2 enrollment
• DD2 Interview data
• Data from clinical examination at enrollment
• Blood and urine samples collected for storage in the DD2 biobank
Additional DD2 data derived from research studies*
Examples
• Analyses of biomarkers and genes from blood and urine samples in the DD2 biobank
• Data from follow-up questionnaires
• Data from intervention studies
Secondary data
From non-centralized data sources
External DD2 data*
Examples
•The Danish Diabetes Database for Adults
•The Danish National Podiatrist Registry
•The Danish National Archive
Individual record linkage on remote research servers
Centralized routine registry data available on remote research servers
Secondary data linked to the DD2 cohort
National Patient Registry
Cause of Death Registry
Cancer Registry
Socio-economic data
Civil Registration System
Laboratory database
Health Services Registry
National Prescription Registry
COVID-19 data
Psychiatric Registry
Medical Birth Registry

Figure 3.1: Overview of data sources in the DD2 cohort (Kristensen et al. (2024)). Notes: *These data can be accessed through collaboration with DD2 research groups.

Chapter 4

Primary data

Primary data can be split into two categories:

4.1 1. DD2 enrollment material

Everything collected during the DD2 enrollment (questionnaire responses, clinical measurements, and biological samples) constitutes the DD2 enrollment material and forms the foundation of DD2 as a distinct data source. This includes:

- DD2 interview data (questionnaire)
- Data from the clinical examination at enrollment
- Blood and urine samples collected for storage in the DD2 biobank in Vejle.

Enrollment in DD2 involves completing the DD2 questionnaire, undergoing the clinical examination, and providing blood and urine samples. The DD2 cohort is defined by individuals who completed the enrollment process. These materials are unique to DD2 and are not available from other sources. All blood and urine samples are stored in the DD2 biobank and may be used for future research, though analysis typically depends on funding from specific studies.

4.2 2. Additional DD2 data

Research studies in the DD2 setup can generate new primary data. The completeness and availability of these primary data depend on when the studies are conducted and when results become available. The data are generated among individuals already enrolled in DD2, and are likely not

collected during the initial enrollment process. This could e.g., be data from studies collecting additional follow-up data (by a follow-up questionnaire or follow-up clinical examinations) or from intervention studies including individuals from the DD2 cohort.
The biological samples (blood and urine) are collected at enrollment and thus considered enrollment material. However, they are stored in the biobank, and can be used for additional analyses in new research studies, e.g., biomarker analyses, genetic studies, or DNA sequencing. These analysis results add to the primary collected data by adding additional baseline information.

Examples of currently available additional DD2 data include

- IDNC questionnaire data, a follow-up questionnaire from 2016 including (follow-up) questions from the initial questionnaire along with multiple questions concerning diabetic neuropathy.
- IDA, DICTA, CsubT2D, etc., are examples of projects initiated within DD2. These projects generate their own additional data and analyses but data are still part of the overall DD2 setup.
- QoL data. After enrollment, an e-mail with a questionnaire about quality of life (QoL) is sent to the individual. Additional follow-up questionnaires are sent after 2 years, 4 years, and 6 years.

These data sources will most likely not be routinely updated with new individuals or additional data, as they are primarily generated as part of specific research projects with closed cohorts. The data can be accessed through collaboration with DD2 research groups.

Chapter 5

Secondary data

Secondary data refer to data from external but non-centralized data sources. They were not initially generated with the intention to be included in DD2. However, as they may be relevant to DD2 research questions, data from these external data sources are linked to the DD2. In practice, this means that data (or a subset of data) are transferred from the external data source to DD2, providing information on the individuals (or a subset of individuals) in DD2. Linkage to external data sources is crucial in DD2, as it allows for information about variables that are resource-intensive to collect in clinical practice and unavailable in the Danish health registries. We currently have data from the following external data sources:

- The Danish Diabetes Database for Adults (DDDA, closed by June 2022)
- The Danish National Podiatrist Registry (fodstatusdatabasen)
- The Danish National Archive (Rigsarkivet), currently including birth data from midwife journals.

In general, updates from these sources may be possible, for example by extending follow-up periods or including additional individuals. The secondary data can be accessed through collaboration with DD2 research groups. Additional agreements may be established with Regionernes Kliniske Kvalitetsudviklingsprogram (RKKP) to enable access to clinical quality data, such as those from the Dansk Diabetes Database (DDiD) and the Dansk kvalitetsdatabase for diabetisk retinopati (DiaBase). Access depends on funding, data availability, and formal collaboration agreements.

Chapter 6

Routine Registry Data

The routine registry data are also secondary data, however, they are only available on remote research servers and not considered part of the DD2 data collection. DD2 research data is uploaded to the servers hosted at Danmarks Statistik (DST, Statistics Denmark) and Sundhedsdatastyrelsen (SDS, The Danish Health Data Authority). On these servers, we have access to Danish medical and administrative registries, such as data from the CPR registry, hospital data from LPR, prescription data, laboratory data etc. Linkage to additional registries and databases is possible after approval of data permission applications sent to Statistics Denmark or the Danish Health Data Authority.

Chapter 7

DD2 questionnaire

Primary DD2 enrollment data

Go to Data documentation

Data from the DD2 questionnaire include the responses from the enrollment, information from the clinical examination along with general information about the individual. The DD2 questionnaire can be downloaded here:

As of summer 2025, DD2 was migrated to a new server setup at Steno Diabetes Center Odense. DCE retains data from the previous setup. Historically, data updates were received regularly, but from the end of 2025, DCE must submit a formal application to obtain new data.

Chapter 8

Data documentation

8.1 DD2core.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (52 x 12,098)	CPR, ProjektID, ydernr	CPR	reg_dato

Data from the DD2 enrollment questionnaire and clinical examination are stored in the `dd2core.sas7bdat` dataset. (Note: The name “dd2core” is not ideal, as there is no definition of what is actually considered “core” data in DD2. Therefore, the data should be referred to as “enrollment data”. The dataset cannot easily be renamed as it is in its current form in the agreements with DST and SDS.)

Data (November 2023) include enrollment information on all N=12,098 individuals enrolled in DD2 by November 2023. Data are on wide format, currently with 52 variables, and include one row per CPR number. A ProjektID is also included for individuals with a blood/urine sample in the biobank. The variable `ydernr` is also included as a identifier and is therefore encrypted by SDS and DST. Individuals are enrolled in DD2 on the enrollment date, `reg_dato`. Data can be updated (new procedure after 2025), and if so, more enrolled individuals are included (same variable set). We have usually received a full new dataset with the current DD2 population - this is to ensure we only have data on individuals we are allowed to have data on, and to ensure any manual corrections in the DD2 setup have been implemented in our data. Data were received as Excel or CSV files and underwent initial data cleaning before being saved in `dd2core.sas7bdat`.

Table 8.2: Illustration of the overall data structure. The dataset is in wide format (52 variables × 12,098 rows), with CPR as the unique key.

Row	CPR	ProjektID	reg_dato	...
1	CPR1	ProjektID1	reg_dato1	...
2	CPR2	ProjektID2	reg_dato2	...
3	CPR3	ProjektID3	reg_dato3	...
...	...	...	...	...
12,098	CPR12098	ProjektID12098	reg_dato12098	...

Chapter 9

DD2 biobank

Primary data derived from DD2 research studies

Go to Data documentation

When individuals are enrolled in the DD2, blood and urine samples are collected and stored in the biobank in Vejle. The samples themselves are considered "primary DD2 enrollment data", and they are all collected at DD2 *baseline*. No automatic or standard analyses are conducted, but DD2 research projects can have the samples analyzed if (additional) analyses are needed - the timing for the analysis results will always be baseline DD2, because it is the time the blood/urine sample was collected. The results from analyses of the blood and urine samples are considered "Additional DD2 data".

For further information about the initial idea about the biobank, please see Christensen et al.
The documentation for the biobank can be downloaded here (Danish):

Chapter 10

Data sources

The data in the biobank are joined from multiple studies and datasets. The individual identifier in the biobank is `ProjektID` and it is unique per CPR number. The project ID should end with the digits -00 (individual). Another variable is `Barcode` which is similar to the project id, but it end with a number (e.g. -12, -19, or -99) and denotes the specific sample for the individual (see documentation above).

Locally at DCE, the (original) data files are stored in the folder:
O:\HE_KEA-DATA-RAW0050\Main Part\data\Input Data Sets\BioBank

Data files from the 2022/2023 rounds are stored in the folder:
O:\HE\KEA-DATA-RAW0050\Biomarkører

Below is a list of the biobank data files stored at DCE (files read in locally). File names might have changed since we initially received them.

10.1 First biomarkers

In the first phase of DD2, a lot of biomarkers were analysed for the first 1,053 individuals enrolled in DD2 (some individuals have later opted out). These individuals are marked by the variable `BlodProve1053patients` (based on the CPR numbers in data file `First1053Patients` (same as in `UrinResultatermedRatio2013Nov`)) and are predominantly enrolled in 2011-2012.

- `First1053Patients.txt`: The file appears to have been received in November 2012. It includes n=1,053 CPR numbers and these individuals are most likely everyone in the database at that time. The file includes results from the initial blood samples. The dataset lacks date information and also unit specifications for the variables hæmolyse, icteri, and lipæmi. The dataset includes the variables:

  - CPR
  - ProjektID
  - C-peptid (n=1,053, pmol/L)
  - GAD (n=1,053, kU/l)
  - Glucose (n=1,050, mmol/L)
  - ALAT (n=1,030, U/L)
  - Hæmolyse (n=1,030)
  - Icteri (n=1,030)
  - Lipæmi (n=1,030)
  - AMYLP (n=1,041, U/L)
  - CRP (n=1,041, mg/L)

We have been told (e-mail from JSN 31.05.2018) not to use the c-peptide measurements from this file (old analysis kit).

Information about some of the variables can be found in Mor et al. (2014).

- `UrinResultatermedRatio2013Nov.txt`: We likely received the data file in December 2013. It includes n=1,053 CPR numbers (same as in `First1053Patients`) with results from the initial urine samples. The dataset includes the variables:

  - CPR
  - Barkode (ends with -19)
  - ALBu_mg_L_ (n=1,041, mg/L, dated November 19, 20, and 23, 2012)
  - KREA_mmol_L_ (n=1,041, mmol/L, dated November 19, 20, and 23, 2012)
  - UPROT_g_L_ (n=1,041, g/L, dated November 19, 20, and 23, 2012)
  - Albumin_Kreatinin_ratio (n=1,041)
  - Kommentar (n=473, <5 with "*PROT-U > 2.5 g/l*" and the rest with "*ALB-U <3 mg/L*")

10.2 Additional biomarkers

- `DataTilDD2medFastende.txt`: We probably received the file in September 2015 (Anne Gedebjerg). It includes n=5,996 CPR numbers with character results on GAD, glucose, and c-peptide, along with a variable about faste. The majority of the individuals in the file are enrolled before 2015, but it is only around 80-85% of all the individuals enrolled before 2016. There are no dates in the data. The dataset includes the variables:

  - Id (ProjectID, ends with -00)

  - GAD (numeric values for n=131 ID numbers, the value "*<0,000*" for n=5,797 ID numbers, "*>525,000*" for n=51 ID numbers, and "*?????*" or "*NA*" for the remaining n=20 ID numbers)
  - Glucose (numeric values for n=4,457 and "*NA*" for the remaining n=1,539 ID numbers)
  - Cpeptid (numeric values for n=5,964 and "*NA*" for the remaining n=32 ID numbers)
  - Fastende ("*Ja*" for n=2,891, "*Nej*" for n=397, and "*Ved ikke*" for the remaining n=2,708 ID numbers) (see below for more information about fasting blood samples)

The majority of the initial 1,053 individuals are also included in this data file. We have been told to use the c-peptide measurements in this data file and not the original data file.

- `WrongCPeptideMeasurements.txt`: The file appears to have been received in November 2015. It includes n=105 ID numbers (end with -00) and the variables cpeptid and NyCpeptid. The values are quite different. We have noted in the syntax that JSN has told us not to use these variables (e-mail 31.05.2018).

10.3 Results from Anne Gedebjerg

- `DD2 resultater.xlsx` (and `DD2 resultater_10_0095`, which is the version where <10 is replaced by 10 and <0,095 by 0,095): We have received the file in September 2017 (from Anne Gedebjerg). It includes n=7,519 barcodes (end with -99) in sheets of 100 or 101 rows, and should include CRP and mannose-binding lectin (MBL) on everyone enrolled by December 2016. See Gedebjerg et al. (2023) for more information. The dataset includes the variables:
  - barkode (ends with -99)
  - CRP (n=7,510, mg/L)
  - MBL (n=7,514, µg/L)

We have been told to keep the CRP measurements from the original file and this file separate. The unit for CRP is mg/L in both data files.

- `Resultater den 250917 Anne Gedebjerg.xlsx`: We have received the file in September 2017 (from Anne Gedebjerg). It includes n=3,116 barcodes (end with -99) and variables regarding MBL expression genotyping (six SNPs in the MBL2 gene). The genotyping was done for the first ~3,000 individuals enrolled in DD2. See Gedebjerg et al. (2020) for more information. The dataset includes the variables:
  - barkode (ends with -99)

  - HL
  - XY
  - PQ
  - 52
  - 54
  - 57
  - HAPLOTYPE

10.4 April 2022 data

During 2022-2023 we received additional data on CRP, c-peptide, and glucose. We had a file with CRP in October 2022, but it is fully included in a file from January 2023 which also includes c-peptide and glucose. The October file is therefore not used during uploads, whereas the January 2023 file has been uploaded to the servers.

- `DD2_cRP_Glucose_Cpep_2022_resultater (1).xlsx`: The data file includes n=3,399 projekt_id. They are all enrolled after the first n=1,053 individuals, but we don't know why they were analysed (not all individuals per year. Maybe in IDA?). The dataset includes the variables: projekt_id, Cpeptid_Barkode, Cpeptid_Resultat, Cpeptid_Måleenhed, Cpeptid_Antal_decimaler, Cpeptid_Dato, Cpeptid_Notat, CRPHS_Barkode, CRPHS_Resultat, CRPHS_Måleenhed, CRPHS_Antal_decimaler, CRPHS_Dato, CRPHS_Notat, Glukose_Barkode, Glukose_Resultat, Glukose_Måleenhed, Glukose_Antal_decimaler, Glukose_Dato, Glukose_Notat.
  - projekt_id (n=3,339)
  - c-peptid (n=2,933, pmol/l, dated 30APR2022 or 01MAY2022)
  - CRP (n=2,478, mg/l, dated 30APR2022 or 01MAY2022)
  - Glucose (n=3,055, mmol/L, dated 02APR2022 or 03APR2022)

10.5 C-peptide and glucose (full)

During the summer 2023 we have received data on c-peptide (July 2023) and glucose (August 2023). These files include **all** measurements from the biobank (cleaned), and results from these files will thus replace all the other measurements from earlier datasets. We now have the files:

- `dd2_all_C_peptide_14July2023.xls`: Includes n=9,762 project_id with data on c-peptide. The file also includes information about analysis date, sampling date, freezing date, unit, and kit. The following data management has been done by DD2 before we received the file:

> Quote from e-mail
>
> ```
> - Alle 1254 observationer, som havde"gamle" målinger (før 28. feb 2015) er erstattet m
> - 6 observationer, som kun havde en "gammel" måling og INGEN opdateret genmåling er sl
> - Der er renset op i data, så hvert individ kun fremgår med én måling, som er analyser
> ```

Please note that c-peptide is now in another unit (it is just the old unit divided by 1000).

- `DD2_glukoser_2023_08_23.xlsx`: The data file includes n=9,563 projekt_id and information about glucose, units, and dates.

10.5.1 HOMA

HOMA values are calculated based on c-peptide and glucose. We use the Oxford calculator which can be downloaded from the website: https://www.dtu.ox.ac.uk/homacalculator/download.php. Since summer 2024, access to the HOMA calculator requires a licence (should be free of charge for “academic researchers”).

HOMA values in the data are calculated based on the c-peptide and glucose measurements received during the summer 2023. HOMA has only been estimated for glucose values in the interval 3.0-25 and c-peptide 0.2-3.5 (because of an updated Oxford HOMA calculator where values out of range cause problems in the excel calculator). Some individuals might therefore have glucose and c-peptide values but no HOMA values in the new data.

10.6 “Pladebiomarkører”

During 2022-2023, we have received new data from additional biomarker analyses (from the Allan Vaag grant). The majority of the new biomarkers are so-called “pladebiomarkører” (as opposed to the current ones which are called “målebiomarkører”) because of the way the analyses are performed. In practice, everyone enrolled as of the day the blood samples were taken from the biobank (in the beginning of 2022) were included in the analysis (approx. the first 9,200 individuals). A small number of individuals have multiple measurements for specific biomarkers, most likely due to sample dilution during the analytical process. Data are in long format and include the following 22 biomarkers (all with unit pg/ml) and the dates refer to when we received the data files:

- TNF-a (April 2022, n=9,202, pg/ml)
- IL-6 (April 2022, n=9,195, pg/ml)
- Ang-Like4 (November 2022, n=9,200, pg/ml)
- FGF-21 (November 2022, n=9,200, pg/ml)

- FGF-23 (November 2022, Hu FGF-23, n=9,200, pg/ml)
- IL1-RA (November 2022, n=9,200, pg/ml)
- Leptin (November 2022, n=9,196, pg/ml)
- RAGE (November 2022, soluble, n=9,200, pg/ml)
- Sclerostin (November 2022, n=9,200, pg/ml)
- U-PAR (November 2022, n=9,200, pg/ml)
- Osteocalcin-1 (February 2023, n=9,203)
- CD163 (April 2023, n=9,047, pg/ml)
- Galectin-3 (April 2023, n=9,008, pg/ml)
- GDF-15 (April 2023, n=9,046, pg/ml)
- NT-proBNP (April 2023, n=9,048, pg/ml)
- Resistin (April 2023, n=9,046, pg/ml)
- Serpin (April 2023, n=9,047, pg/ml)
- YKL-40 (April 2023, n=9,045, pg/ml)
- Osteopontin (June 2023, n=9,204, pg/ml)
- Adiponectin (July 2023, n=9,204, pg/ml)
- Follistatin (July 2023, n=9,204, pg/ml)
- MPO (July 2023, n=9,204, pg/ml)

An overview of the biomarkers (table from the Allan Vaag grant application) can be found here:

An additional document combining overview sheets, method descriptions, and quality logs from some of the analysis rounds can be downloaded here:

The data files with "pladebiomarkører" and "målebiomarkører" don't have the same format (e.g., long vs. wide format) and we have thus not combined the data files.

10.6.1 Data files (pladebiomarkører)

In the following section we will go through the data files we have received.

- April 2022, `220211 Vplex_final.xlsx` with 2 sheets (data and background information). Data received in April 2022 but probably analyzed in February 2022 (no dates in data, but based on date stamps in file names). The data file includes data from n=9,294 DD2 individuals on IL-6 and TNF-a. The first sheet, `Vplex sample results_final`, includes the variables: Sample (id, ends with -12), Sample_Group (=Sample in all rows), Assay (either TNF-a or IL-6), Calc___Conc___Mean (results), RANGE (value 1 or 2), Plate_Name (each Plate_Name is used 78 times).
  - TNF-a (n=9,202, pg/ml)
  - IL-6 (n=9,195, pg/ml)

  The second sheet, `Vplex complete final`, includes background information (rådata) about the sample from the sample_groups *Sample* (n=9,024), *Standards* (n=1,888), and *Internal Control* (n=236). The sheet includes the variables Plate_Name, Sample_Group, Sample, Assay, Well, Signal, Mean, CV Calc___Concentration, Calc___Conc___Mean, Calc___Conc___CV, ___Recovery, ___Recovery_Mean, Detection_Limits___Calc___Low, Detection_Limits___Calc___High, Detection_Range, Detection_Range_yesno, Quantification_range, Quantification_range_yesno, RANGE.

We also received the data file `DD2 quality log_panel 1 edited_TNF IL6.xlsx` but is has not been used.

- November 2022, `8plex data final.xlsx`, with 9 sheets (overview + 8 biomarkers). We received the data file in November 2022. There are no date stamps indicating when the analyses were performed. The data file include information on n=9,204 individuals (based on sample ID ending with -12). For each assay, the data file includes the variables sample (ID), assay, calc___conc___mean (result), RANGE, and plate_name. In an e-mail, we have been told the unit is pg/ml for all assays.

  - Ang-Like4 (n=9,200, pg/ml)
  - FGF-21 (n=9,200, pg/ml)
  - FGF-23 (Hu FGF-23, n=9,200, pg/ml)
  - IL1-RA (n=9,200, pg/ml)
  - Leptin (n=9,196, pg/ml)
  - RAGE (soluble, n=9,200, pg/ml)
  - Sclerostin (n=9,200, pg/ml)
  - U-PAR (n=9,200, pg/ml)

  We also received data files `eight biomarkers with CPR.xlsx` and `seven biomarkers with CPR.xls` but these have not been used.

- February 2023, `DD2 osteocalcin final.xlsx` with 3 sheets (overview, results, and *rådata*). We received the file in February 2023, but we have no indication of when the analyses were performed. The file includes n=9,204 individuals (sample, ends with -12). The data file includes the variables sample (ID), assay, calc___conc___mean (result), RANGE, and plate_name.

  - Osteocalcin-1 (n=9,203)

  The sheet *rådata* includes detailed information about each of the plates.

- April 2023, `DD2_7plex_data_final.xlsx` with 9 sheets (overview + 7 biomarkers + additional sheet with sample names). We received the data file in April 2023. There are no date stamps indicating when the analyses were performed. The data file include information on n=9,048 individuals (based on sample ID ending with -12). For each assay, the data file includes the variables sample (ID), assay, calc_conc_mean (result), RANGE, and plate_name. In an e-mail, we have been told the unit is pg/ml for all assays.

  - CD163 (n=9,047, pg/ml)
  - Galectin-3 (n=9,008, pg/ml)
  - GDF-15 (n=9,046, pg/ml)
  - NT-proBNP (n=9,048, pg/ml)
  - Resistin (n=9,046, pg/ml)
  - Serpin (n=9,047, pg/ml)
  - YKL-40 (n=9,045, pg/ml)

- June 2023, `Osteopontin data final (1).xlsx` with 3 sheets (overview + data + *rådata*). We received the file in June 2023, but we have no indication of when the analyses were performed. It includes n=9,207 ID numbers (end with -12, plus a note that it means "EDTA plasma fraction") with data on osteopontin.

  - Osteopontin (n=9,204, pg/ml)

- July 2023 (1), `DD2_adiponectin_blue panel_final (1).xlsx` with 3 sheets (overview + data + *rådata*). We received the file in July 2023, but we have no indication of when the analyses were performed. It includes n=9,204 ID numbers (sample, end with -12) with data on adiponectin.

  - Adiponectin (n=9,204, pg/ml)

  The sheet *rådata* includes detailed information about each of the plates.

- July 2023 (2), `DD2 red panel_final.xlsx` with 4 sheets (overview + data (Follistatin + MPO) + *rådata*). We received the file in July 2023, but we have no indication of when the analyses were performed. It includes n=9,204 ID numbers (sample, end with -12) with data on follistatin and MPO

  - Follistatin (n=9,204, pg/ml)
  - MPO (n=9,204, pg/ml)

  The sheet *rådata* includes detailed information about each of the plates.

Chapter 11

Fasting

Is the individual fasting? A simple question, yet, difficult to assess.

Data from the DD2 questionnaire itself include the variable Er_patienten_fastende_ (whether the patient is fasting). Currently, around 75% of the individuals have answered that they are fasting. This variable can be used on its own, but should probably be combined with the variable Tages_der_blodproeve_i_forbindel (whether the blood sample was taken at the same time as the questionnaire was answered). If the fasting patients are restricted to include only those whose blood sample was taken at the same time as the questionnaire was answered, then around 72% of the individuals are defined to be fasting.

The data file DataTilDD2medFastende.txt received in September 2015 included information on fasting state for n=5,996 individuals ("*Ja*" for n=2,891, "*Nej*" for n=397, and "*Ved ikke*" for n=2,708 individuals). This file will not be updated and we therefore don't get new information on this fasting variable. We don't know how this variable was defined, but it is probably based on information on the blood sample itself. By adding the information from this file (variable name NewFastende) to the variable Er_patienten_fastende_, an additional 447 individuals can be defined as fasting (410 with missing information in Er_patienten_fastende_ and 37 who replied not to be fasting in Er_patienten_fastende_).

Currently, the fasting state has been defined by the following (SAS) algorithm combining all the files and stating that the individual is fasting if we have any indication that this could be the case (macro: AdditionalVars_Faste):

```
    if Tages_der_blodproeve_i_forbindel in: ('Ja') then do;
        if Er_patienten_fastende_=:'Ja' or NewFastende=:'Ja' then Faste='Ja';
        else if Er_patienten_fastende_=:'Nej' or NewFastende=:'Nej' then Faste='Nej';
    end;
    else if Tages_der_blodproeve_i_forbindel in: ('Nej') then do;
        if NewFastende=:'Ja' then Faste='Ja';
        else if NewFastende=:'Nej' then Faste='Nej';
    end;
    else if Tages_der_blodproeve_i_forbindel in (' ') then do;
        if Er_patienten_fastende_=:'Ja' or NewFastende=:'Ja' then Faste='Ja';
        else if  Er_patienten_fastende_=:'Nej' or NewFastende=:'Nej' then Faste
    end;
```

Note: DD2 plan to make an “official” definition of Faste.

Upon enrollment, the individuals are informed to be fasting: no food/liquid (except water) from 10.00 o’clock the night before. Also, while fasting, the patient should not take any glucose-regulating drugs.

Chapter 12

Data documentation

12.1 biobank.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (48 x 11,381)	CPR, ProjektID	CPR (ProjektID)	VejleDato

The biobankdataset include analysis results from "målebiomarkører". In practice, it includes results from all analyses not part of the 22 variables from the "pladebiomarkører" in the Allan Vaag grant. A few variables from dd2core (e.g. reg_dato and Er_patienten_fastende_) are included in the biobank dataset.

Data include analysis results from successful analyses. Not all analyses are performed for all individuals (missing data), and we don't have more information about specific analyses (i.e., project, analysis method, unit/kit, non-successful analyses etc.). In some versions of the dataset, rows are included for all CPR numbers in the population, even if no analysis results are available.

Table 12.2: Illustration of the overall data structure. The dataset is in wide format (48 x 11,381), with CPR or ProjektID as the unique key.

Row	CPR	ProjektID	Analysis1	Analysis2	Analysis3	...
1	CPR1	ProjektID1	num.	num.		...
2	CPR2	ProjektID2	num.		num.	...
3	CPR3	ProjektID3				...
4	CPR4	ProjektID4	num.	num.	num.	...
...	...	...	...	...	...	...

Row	CPR	ProjektID	Analysis1	Analysis2	Analysis3	...
12,098	CPR12098	ProjektID12098	num.	num.	num.	...

12.2 biomark.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Long (9 x 201,243)	CPR, ProjektID, ydernr	CPR*Assay	(Vejledato)

The `biomark`dataset include analysis results from the 22 “pladebiomarkører” in the Allan Vaag grant. No dates are included in the dataset, however, analyses are performed on the enrollment blood sample. The dataset include approximately 9,200*22=202,400 rows. In principle, CPR*Assay should be the unique key, however, some analyses are performed multiple times per individual (due to dilution in the analysis).

Table 12.4: Illustration of the overall data structure. The dataset is in long format (9 x 201,243), with CPR*Assay as the unique key.

Row	CPR	ProjektID	Assay	Value	Info	...
1	CPR1	ProjektID1	TNF-a	num.	...	...
2	CPR1	ProjektID1	IL-6	num.	...	...
3	CPR1	ProjektID1	Ang-Like4	num.	...	...
...	...	...	...	...	...	...
22	CPR1	ProjektID1	MPO	num.	...	...
23	CPR2	ProjektID2	TNF-a	num.	...	...
24	CPR2	ProjektID2	IL-6	num.	...	...
...	...	...	...	...	...	...

Chapter 13

Genetic variables

Primary data derived from DD2 research studies

Go to Data documentation

We received data on genetics/polygenic risk scores (PRS) in multiple rounds. In the following section we will go through the data files we received. Of note, we also received PRS data in October 2023 which was used for one of the birth weight studies (Hansen et al. (2023)), but these data have not been uploaded to the servers.

On SDS, the first and second rounds of data were originally included in the `IL6_TNF` data, but this data source has now been split into `Biomark` including the 22 "pladebiomarkør" biomarkers and `Genetik` including the genetics data described here. The variables in `Genetik` were renamed, i.e., `Assay = GeneMarker`, `Calc_Conc_Mean = Value`, and `CV_on_plate = data_round`.

13.1 Round 1 (April 2024 - Gen1)

Data include 9,997 observations and 104 variables. Some individuals had missing ID (but potentially non-missing barcode) and some had duplicate barcodes. The n=215 individuals with missing ID were checked by JSN, and 85 of these were denoted "ok". For around 500 individuals, information on all the PRS values were deleted in the genetic analysis (see README file for specific reasons). Data were restricted to n=9,856 individuals with valid, unique, and non-missing ID and barcode, and transformed to long format. The variables geno_id, rem, p_id (text), DD2_projectid, DD2_barcode were removed from data (to align with the format from the original biobank file, `IL6_TNF`). Before upload to SDS (`IL6_TNF_240411`), 69 individuals were excluded as they were not among the n=11,381 individuals in the DD2 cohort (SDS, April 2023). Data were marked with `data_round="Gen1"`.

The PRS `oram_T1D` has been replaced by `oram_T1D_v2` in Round 2 as there was an error in the calculation from Round 1. The variable from Round 1 should therefore not be used.

13.2 Round 2 (July 2024 - Gen2)

Data include 9,997 observations and 28 variables. Using the same procedure as in Round 1, we end up with n=9,856 individuals with valid, unique, and non-missing ID and barcode. Again, data were transformed to long format, and the variables geno_id, rem, p_id (text), DD2_projectid, DD2_barcode were removed. Before upload to SDS (`IL6_TNF_240711`), 69 individuals were excluded as they were not among the n=11,381 individuals in the DD2 cohort (SDS, April 2023). Data were marked with `data_round="Gen2"`.

13.3 Round 3 (December 2024 - Gen3)

Data include 9,997 observations and 65 variables. Using the same procedure as in Round 1 and Round 2, we end up with n=9,856 individuals with valid, unique, and non-missing ID and barcode. Data were marked with `data_round="Gen3"`.

For single SNPs, the original variable name corresponded to `chr:position:allele1:allele2_e` however, SAS replaces `:` with `_` and adds `_` in the beginning of the variable name (as a variable name can't start with a number), i.e., the variable name is currently `_chr_position_allele1_allele2_effect-allele`. Additionally, a document including the r2 values is available on the servers.

13.3.1 Round 3, additional data for Tarun (Tar1)

In addition to the Round 3 genetics data, we received a data file including a broader list with around 1300 SNPs (Tarun Veer Singh Ahluwalia is lead on the project). The file thus included 9,997 observations and 1309 variables, and we again restricted to the n=9,856 individuals with valid, unique, and non-missing ID and barcode. Data were marked with `data_round="Tar1"`, and variable names correspond to `_chr_position_allele1_allele2_effect-allele`. A document including the r2 values is available on the servers.

Chapter 14

Data documentation

14.1 genetik.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Long (8 x 14,572,843)	CPR, ProjektID	CPR*GeneMarker-	

The genetik dataset includes analysis results from the three data rounds. There are no dates in the dataset, but all analyses are performed on the enrollment samples. Data can be stratified by round using the variable data_round with values Gen1, Gen2, Gen3, and Tar1. A total of 1,489 different gene markers (or PRS or whatever they are) are currently included in the data. Data include 9,787*1,489=14,572,843 rows.

Table 14.2: Illustration of the overall data structure. The dataset is in long format (8 x 14,572,843), with CPR*GeneMarker as the unique key.

Row	CPR	ProjektID	data_round	GeneMarker	Value	...
1	CPR1	ProjektID1	Gen1	ACSL1_rs221	num.	...
2	CPR1	ProjektID1	Gen1	ACSL1_rs624	num.	...
4	CPR1	ProjektID1	Gen1	ACSL1_rs756	num.	...
...	...	...	...	...	...	...
1,489	CPR1	ProjektID1	Tar1	wang_ST	num.	...
1,490	CPR2	ProjektID2	Gen1	ACSL1_rs221	num.	...
1,491	CPR2	ProjektID2	Gen1	ACSL1_rs624	num.	...
...	...	...	...	...	...	...

Additional documentation and README files (from Mette Andersen, KU) are available upon request. Among others, they include information about GeneMarker definition.

Chapter 15

Neuropathy questionnaire (IDNC)

Primary data derived from DD2 research studies

Go to Data documentation

IDNC is short for “International Diabetic Neuropathy Consortium”. Read more about it here: https://idnc.au.dk/about-idnc.

A list of 7,011 CPR-numbers from all individuals enrolled in DD2 by February were sent to the CPR registry to get their addresses. The only criteria to be included in the questionnaire study were that individuals should be enrolled in DD2 (February 2016), be alive (24 May 2016), and have a valid Danish address (24 May 2016). A total of 6,726 questionnaires were sent out in June 2016 (07 June 2016), and a reminder was sent in September 2016 (12 September 2016) and again in October 2016 (10 October 2016) to those who had not provided a response. A total of 5,755 questionnaires were returned during the fall of 2016 and the inclusion ended in January 2017 (24 January 2017). The questionnaire included questions about height, weight, smoking and alcohol, and physical activity. These questions were intended to be follow-up questions similar to those from the baseline questionnaire. In addition, the questionnaire included questions about falls, quality of life, sleep, mental health, neuropathy (in hand and feet) and pain. The questionnaire thus included the Michigan Neuropathy Screening Instrument questionnaire (MNSIq) and the Douleur Neuropathique en 4 Questions (DN4).

Read more about the questionnaire in Gylfadottir et al. (2020) and Christensen et al. (2020).

The questionnaire (in Danish) can be downloaded here:
An English translation of the questionnaire is included in the supplement to Gylfadottir et al. (2020).

The questions and variable definitions for the original variables can be downloaded here:

Chapter 16

Data documentation

16.1 IDNC.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (162 x 5,658)	CPR	CPR	reg_dato, IDNCData_date

The `IDNCData_date` takes the value `01JUN2016` for all individuals, as this is the date of the first IDNC data round. The local data also include the variable `tid` with the values `T1`, `T2` and `T3`, indicating in which round the individual answered the questionnaire (based on whether or not a reminder was sent). This variable is not included in the datasets on the servers.

Table 16.2: Illustration of the overall data structure. The dataset is in wide format (162 x 5,658), with CPR as the unique key.

Row	CPR	IDNCData_date	reg_dato	...
1	CPR1	01JUN2016	reg_dato1	...
2	CPR2	01JUN2016	reg_dato2	...
3	CPR3	01JUN2016	reg_dato3	...
...	...	...	...	...
12,098	CPR12098	01JUN2016	reg_dato12098	...

The original data are stored locally at DCE and a slightly modified version (IDNC.sas7bdat) has been uploaded to DST and SDS. The changes are mainly due to similar definitions of variables, variable names and labels, and to comply with the SDS/DST regulations.

16.1.1 Variables

The original questionnaire data included a total of around 109 variables (including some variables from the data cleaning process added by the working group). To ease working with data and to have some pre-defined variables/definitions, an additional 78 variables were created by the working group.

The SAS syntax for the definition of the additional variables is available here:

And the corresponding SAS formats for the additional variables are available here:

(SAS syntax files can be opened using notepad if SAS is not installed)

Chapter 17

DDDA

Secondary data from non-centralized data source

Go to Data documentation

Danish Diabetes Database for Adults (DDDA) (also known as Dansk Voksen Diabetes Database (DVDD) or Det Nationale Indikatorprojekt (NIP) Diabetes) is a no longer active database from regionernes kliniske kvalitetsudviklingsprogram (RKKP). It was closed by 30 June 2022. It has now been replaced by Dansk Diabetes Database (DDD/DDiD), which has been active since 1 July 2022. DDiD currently mainly uses registry data, however, as additional data are still being recorded in the clinic, it is worth checking the database in the future, as additional variables might be included (e.g. smoking, blood pressure, BMI, etc.).

Extracts from the online DDDA documentation (downloaded 07 September 2023) about the population and variables can be downloaded here:

And additional documents:

See also Jørgensen et al. (2016) for further information about the database.

Chapter 18

Data

At DCE, we have data from DDDA in the period from May 2005 to January 2022 based on the n=10,241 individuals enrolled in DD2 per January 2022. As the database is now closed, this is the final data set and we will not get any updates of it.

In the latest version of the DDDA data, around 80% (n=8,512) of the individuals in DD2 had at least one record in DDDA. The records in DDDA are uniquely defined based on the CPR-number and the variable `status_dato` which is the date that defines the data entry. There is a median of 3 (IQR 2-5) records per individual, 85% had at least one record prior to enrollment in DD2, and there was a median of -56 days (IQR -413-294) from the enrollment date to the nearest `status_dato` (negative value meaning that `status_dato` was before enrollment in DD2). The `status_dato` is not necessarily the date of the examination/results/visit. Many of the variables thus have an associated date variable that is different from `status_dato`.

When using data from DDDA, it is important to keep in mind the missing data. Data may be missing either because it was not reported or because the individual is simply not registered in DDDA. Thus, when reporting data from DDDA, e.g., in a baseline table, remember to consider what "100%" should represent, as the interpretation of the result varies; is it x% of the entire study population or x% of the study population also registered in DDDA.

In addition, the selection and timing of data should be considered. Individuals registered in DDDA fulfill the criteria for this database (see population documentation above) which might induce some selection bias. Individuals in DDDA might not be as newly diagnosed as individuals in DD2. The records in DDDA are defined based on `status_dato`, which is "den dato, hvor der årligt, i relation til databasen, gøres status over hvornår personer senest har fået foretaget forskellige relevante undersøgelser" (see website). The dates listed for each of the examinations might be long after (or before) enrollment in DD2, and for each scientific study question, it is thus crucial to consider in which time period data can be considered valid. An example could be to only include measurements in the period from 5 years prior to enrollment to 6 months after enrollment and then among these, include the one closest to the enrollment date as a "baseline" value. This depends on the specific study.

Chapter 19

Data documentation

19.1 DDDA.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (75 x 30,806)	CPR	CPR*status_dato	status_dato, specific date variables

N=8,512 distinct CPR numbers with non-missing data are included in the data, i.e., individuals with data in DDDA. A total of N=10,219 individuals are included (merged with DD2core). A record is unique based on CPR and status_dato, however, for many of the variables, it is worth also considering the corresponding var_dato.

Table 19.2: Illustration of the overall data structure. The dataset is in wide format (75 variables × 30,806 rows), with CPR*status_dato as the unique key.

Row	CPR	status_dato	Var	Var_dato	...
1	CPR1	status_dato1.1	num.	Var_dato1.1	...
2	CPR1	status_dato1.2	num.	Var_dato1.2	...
3	CPR1	status_dato1.3	num.	Var_dato1.2	...
...	...	...	...	...	...
10	CPR1	status_dato1.10	(missing, no record)	(missing, no record)	...
11	CPR2	status_dato2.1	num.	Var_dato2.1	...
12	CPR2	status_dato2.2	num.	Var_dato2.2	...

Row	CPR	status_dato	Var	Var_dato	...
...	...	...	...	...	...
30,806	CPR10219	(missing, not in DDDA)	(missing, not in DDDA)	(missing, not in DDDA)	...

19.1.1 Variables from previous versions

In the current DDDA data, some variables from the database were not delivered. It was decided that the best suitable solution was to include these variables from older versions of data (and thus missing for some new individuals).

- From the 2021 data: `shared_care` and `Plasmakreatinin_operator`
- From the 2018 data (also missing in the 2021 delivery): `diag_dato`, `HbA1c_kode`, `HDLcholesterol`, and `HDLcholesterol_undersoegelse_kod`. These variables are included with a “_2018” postfix.

Chapter 20

Podiatrist data

Secondary data from non-centralized data source

Go to Data documentation

The DD2 cohort is linked to The Danish National Podiatrist Registry (the Podiatrist Database, fodstatusdatabasen). This database was established in 2011, and includes information on examinations performed by Danish podiatrists working under a collective agreement with the Danish Administrative Regions (Kristensen et al. (2024)). Records include symptoms of neuropathy and information from a clinical examination of the lower extremities (including circulatory and neuropathy status).

Individuals in DD2 enrolled between 2011 and 2022 are linked to the Podiatrist Database. Approximately 60% of the DD2 participants have at least one record in the database, and they have median 4 (IQR 2-6) records. DD2 participants without records may have had foot examinations by other healthcare professionals (e.g., diabetes outpatient clinics, general practitioners), a private podiatrist not working under the collective agreement or may not have been examined at all (Kristensen et al. (2024)).

Chapter 21

Data documentation

21.1 fodstatus.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (88 x 29,214)	CPR	CPR*statusdate	statusdate

The dataset include podiatrist data for a total of N=6,734 individuals. Data are uniquely identified via CPR*statusdate.

Table 21.2: Illustration of the overall data structure. The dataset is in wide format (88 variables x 29,214 observations), with CPR*statusdate as the unique key.

Row	CPR	statusdate	Variable1	Variable2	...
1	CPR1	statusdate1.1	num.	char.	...
2	CPR1	statusdate1.2	num.	char.	...
...	...	...	...	...	...
5	CPR1	statusdate1.5	num.	char.	...
6	CPR2	statusdate2.1	num.	char.	...
7	CPR2	statusdate2.2	num.	char.	...
...	...	...	...	...	...
29,214	CPR6734	statusdate6734.1	num.	char.	...

Chapter 22

Birth data

Secondary data from non-centralized data source

Go to Data documentation

Data from Danish midwife journals have been collected for all individuals enrolled in DD2 per 22 February 2022, *i.e.,* a total of n=9,549 individuals. Data include information on birth weight, birth length, born at term (yes/no, and if no then also the number of weeks), and twin (yes/no).

The afleverings-/afslutningsrapport from Rigsarkivet (the Danish National Archive) can be downloaded here (Danish):
There is similar description in English in the electronic supplemental material from Hansen et al. (2023):

CPR-numbers for a total of n=9,549 individuals were sent to Rigsarkivet and n=8,896 fulfilled the criteria for potentially having available birth information (born in the period from ~1920 to 1988, born in Denmark, and identifiable through the biological mother's name). During the retrieval, data were uniformized (*e.g.,* to include birth weight in grams and not pounds) and proofread.

Data include n=9,549 CPR-numbers (one CPR-number is included twice), n=8,346 have non-missing birth weight, and a total of n=8,364 had a non-missing value in at least one of the five birth variables. A total of 8,869-8,364=532 were not included despite fulfilling the criteria (not registered[1]).

[1] It is not marked in the data whether missing was due to not being eligible or due to other reasons

Chapter 23

Control population

In addition to the individuals from DD2, a control population was recorded. For every individual in DD2, approximately two random individuals were selected based on the births from the same midwife information sheet (with birth information on ~6–8 different births). Controls from the same midwife record thus served as a match on date of birth, midwife, and geographical location. A total of n=18,210 individuals are recorded in the control population and data include information on year and month of birth, birth weight, birth length, sex, born at term, twin, and geographical location (lægekreds/fødselsamt). No CPR-numbers are available for the control population (data permissions and ethical and practical reasons) and neither are links between the controls and the individuals in DD2. Data on controls are currently not available on the servers[1].

Selected pages from the opgavebeskrivelse (del 1 and del 2) can be downloaded here (Danish):

[1] The current data permission includes only individuals in DD2 and therefore not the matched controls. As CPR-numbers are not available and controls ultimately cannot be linked to their respective DD2 individual there is no reason to upload data to the servers because all analyses can be performed locally.

Chapter 24

Data documentation

24.1 foedselsdata.sas7bdat

Format (var x obs)	Id variables	Unique key	Important dates
Wide (9 x 9,544)	CPR	CPR	-

The dataset include birth data for N=9,544 individuals identified via CPR. There are no dates in the dataset, however, information relates to birth date.

Table 24.2: Illustration of the overall data structure. The dataset is in wide format (9 variables x 9,544 observations), with CPR as the unique key.

Row	CPR	foedselsvaegt	foedselslaengde	fuldbaarenhed	uger_for_tid	tvillingefoedsel
1	CPR1	num.	num.	Ja		0
2	CPR2	num.	num.	Nej	2 til 3	0
3	CPR3	num.	num.	Ja		0
...	...	...	...	...	...	...
9,544	CPR9544	num.	num.	Ja		1

Chapter 25

SDS

Secondary data from centralized routine registry data available on remote research server

25.1 Sundhedsdatastyrelsen (SDS)/The Danish Health Data Authority

Sundhedsdatastyrelsen - FSEID-00006014.

Download the "udtræksbeskrivelse" here:
Dataspecifikation
Dataoversigt

25.2 Population

The DD2 data are uploaded to SDS, and they check the validity of the CPR-numbers before encrypting them. The dataset `DS_EXT_DD2_POP` includes all individuals with a valid CPR-number. Individuals with a non-valid CPR-number are still included in the uploaded data, e.g. in `DS_EXT_DD2CORE`, but they will have a missing encrypted CPR-number (`CPR_ENCRYPTED`) and can thus not be linked across data sources. Previously (when we had local data), we would notify DD2 (Odense) about non-valid CPR-numbers, but this is no longer possible as we are not allowed to identify the individuals based on individual-level data on the servers.

The data files uploaded by us have the prefix `DS_EXT_`. Data are available in 2 versions; the initial upload version (without suffix) and the updated versions (with suffix _230601, _231026, etc., corresponding to the date of the update on SDS). The registry data (views) will always be based on the newest population, i.e., individuals in the newest version of `DS_EXT_DD2_POP`.

25.3 Rawdata available on SDS

Data on SDS are based on data views, i.e., we have access to data updated (almost) on a day-to-day basis. The update file (SAS `IN06014.OpdateringsOversigt`) contains the time of the current update for each dataset in the library. The metadata file (SAS `metadata.metadata_allcolumns`) contains information about variables in the data. For research projects, in order to be able to reproduce results, remember to save a copy of the research data in a suitable project folder on SDS.

> **Note**
>
> Using SAS, if you extract data from views you will likely save a lot of time by writing code that is proper sql, rather than using SAS-specific syntax

```
Example:
1)

proc sql;
  create table x as
  select *
  from data.lab_lab_dm_forsker(where = (analysiscode in: ([list of npucodes]))
quit; * takes ~40 minutes to run

2)
proc sql;
  create table x as
  select *
  from data.lab_lab_dm_forsker
  where analysiscode eqt [npucode1] or ... or analysiscode eqt [npucodeN];
quit; * takes ~40 minutes to run


3)
proc sql;
  create table x as
  select *
  from data.lab_lab_dm_forsker
  where analysiscode like [npucode1%] or ... or analysiscode like [npucodeN%];
```

```
quit; * takes ~10-20 seconds to run
```

Further documentation on the main data sources can be found here:

- CPR-registret (CPR): documentation
  References: Schmidt et al. (2014)
- Landspatientregistret (LPR2 and LPR3): documentation, LPR3, LPR3 vejledning
  References: Schmidt et al. (2015)
- Lægemiddelstatistikregistret (LRS/LMS): documentation
  References: Kildemoes et al. (2011), Johannesdottir et al. (2012), Pottegård et al. (2017), Ehrenstein et al. (2010)
- Laboratoriedatabasen (LAB-F/RLRR): documentation
  References: Grann et al. (2011), Arendt et al. (2020)

25.3.1 Access to data

All DD2-researchers with access to FSEID-00006014 can (in principle) work with all the datasets. As these datasets are considered rawdata, they are cannot be edited/changed/moved/renamed/etc. To access data, please do the following:

- **SAS**: The data are already available in the libraries `in06014` and `metadata`
- **STATA**: Run the (relevant parts of the) DO file located at `F:\Projekter\FSEID00006014\DB Connections FSED00006014.do`
- **R**: Run the (relevant parts of the) R file located at `F:\Projekter\FSEID00006014\DB Connections FSED00006014.R`

In each research project, it is the individual researchers' responsibility to ensure data usage is allowed.

25.4 Data uploads

The first data were uploaded in 2022 when the server access was initially established. Since then, the data have been updated multiple times.

Short description	Detailed description	Date	Affected datasets
Original upload	Upload of datasets for the initial server access	Summer 2022	DD2CORE, DD2_POP, BIOBANK, DDDA, FOEDSELSDATA, IDNC, IL6_TNF
General update	Population update and inclusion of data source FODSTATUS. The biobank was updated with additional measurements of CRP, c-peptide, glucose, and HOMA (from April 2022), and additional "pladebiomarkører" were added to IL6_TNF. DDDA data were updated (June 2022) before the registry closed. IDNC was trimmed to the current population.	June 2023 (_230621)	DD2CORE, DD2_POP, BIOBANK, DDDA, FOEDSELSDATA, IDNC, IL6_TNF, FODSTATUS

Short description	Detailed description	Date	Affected datasets
Upload of biobank data	C-peptide and glucose (and therefore also HOMA) were changed, and the remaining "plade-biomarkører" were added to IL6_TNF, which is now complete	Fall 2023 (_231026)	BIOBANK, IL6_TNF
Upload of biobank data (genetics)	Genetic variables (polygenic risk scores) were added to IL6_TNF (long format). All new variables are labelled "Gen1"	April 2024 (_240429)	IL6_TNF
Upload of biobank data (genetics) and DICTA variable	Additional genetic variables were added to IL6_TNF. In DD2CORE the DICTA variable was added.	July 2024 (_240711)	DD2CORE, IL6_TNF

Short description	Detailed description	Date	Affected datasets
Population update and upload of genetics	The population (and registries) was updated (November 2023) and additional genetic variables were included. We made the SDS update in the same round where the DST data were updated (genindstilling via DDV), and decided to rename IL6_TNF to BIOMARK including only the "plade-biomarkører", and including the genetic variables in a separate dataset called GENETIK. The dataset INFODATA including e.g., DICTA was also included.	December 2024 (upload not available as we are waiting for SDS updates)	DD2CORE, DD2_POP, BIOMARK, GENETIK, INFODATA
December 2024 upload	The upload from December 2024 was made available and we added CAR and DAR from the "omlægning af registre" (still waiting for LAB).	June 2025	DD2CORE, DD2_POP, BIOMARK, GENETIK, INFODATA (and CAR+DAR from registries)

NB: we try to distinguish between "update" and "upload". An "update" means that the population, and therefore also the registry data, has changed, and that the entire DD2 cohort should be seen as a new version. By "upload" we mean additional data uploaded to the current population. This could for example be adding analyses from the biobank or adding a new data source, i.e., something that doesn't affect the population itself. The plan is to make one annual update (November/end of the year) and two to three additional uploads during per year.

Chapter 26

DST

Secondary data from centralized routine registry data available on remote research server

26.1 Danmarks Statistik (DST)/Statistics Denmark

DST Projektnr. 708637 (and FSEID-00006159 transfer from SDS). The project is placed under the “Projektdatabaseordning (705004)”. The DD2 projects itself had the project number 708637 (via 705004), and the agreement with DST makes registry data available. Data from Det Psykiatriske Centrale Forskningsregister (PCR), Laboratoriedatabasens Forskertabel (LAB-F), Overvågningsdata - COVID-19-testsvar fra SSI (OVD_SSI), Det Danske Vaccinationsregister - COVID-19-vaccinationsdata fra SSI (DDV_SSI), and Cancerregisteret (CAR) are tranferred from SDS (FSEID-00006159). Data from Lægemiddelstatistikregisteret/Lægemiddeldatabasen (LSR/LMDB) is included via DDV (DST), but based on the agreement with SDS (FSEID-00006159).

> **!** Important
>
> Tjek om aftale + bilag må ligge offentlig tilgængeligt.

Download the general agreement including the agreement about medicinal data here:
Aftale (mail)
Bilag LMDB

26.2 Population

The DD2 data are uploaded to DST, and the population include all individuals in DD2CORE (encrypted CPR). Registry data (Raw data -> Grunddata) is linked, and based on the individuals in the dataset `pniveau`.

26.2.1 Formats

On DST formats from DST and IDNC should be included manually. Here is an example of how it can be done in SAS

```
libname fmt "E:\Formater\SAS formater i DAnmarks Statistik\FORMATKATALOG" acces
options fmtsearch=(fmt.disced fmt.times_personstatistik fmt.dream fmt.statistik

libname idnc "path" access=readonly;
options fmtsearch=(idnc.idnc_formats);
```

26.3 Data uploads

The first data were uploaded in 2022 when the server access was initially established. Since then, DST has changed their system to Danmarks Datavindue (DDV). In November 2024, we had to make a "genindtilling" via DDV and in principle make a new data agreement. We used this opportunity to make some changes in the data structure (e.g., from IL6_TNF to BIOMARK and GENETIK)

Short description	Detailed description	Date	Affected datasets
Original upload	Upload of datasets for the initial server access. Data from the original upload are located in the folder with "Eksterne data" under the rawdata folder (Rawdata > 708637 > Eksterne data). DD2 data are in the 20230104_FraSDS folder, whereas SDS data are in the folder 20221020_FraSDS. Registry data are in the Grunddata folder	Summer 2022	DD2CORE, DD2_POP, BIOBANK, DDDA, FOEDSELSDATA, IDNC, IL6_TNF

Short description	Detailed description	Date	Affected datasets
General update (DDV)	Population update (November 2023), updated biobank data (variables from April 2022), including the new data source FODSTATUS, and a general update of registry data. IL6_TNF was updated with all 22 "plade-biomarkører" and renamed to BIOMARK. The dataset GENETIK was also included. The dataset INFODATA including e.g., DICTA was also included.	November 2024/March 2025 (upload available in the folders 20250402_FraSDS and 03032025). We are still waiting for registry updates (mainly due to changes at SDS)	DD2CORE, BIOBANK, DDDA, FOEDSELSDATA, IDNC, FODSTATUS, BIOMARK, GENETIK, INFODATA.

NB: we try to distinguish between "update" and "upload". An "update" means that the population, and therefore also the registry data, has changed, and that the entire DD2 cohort should be seen as a new version. By "upload" we mean additional data uploaded to the current population. This could for example be adding analyses from the biobank or adding a new data source, i.e., something that doesn't affect the population itself. The plan is to make one annual update (November/end of the year) and two to three additional uploads during per year.

Registry data on the forskermaskine is updated regularly. LPR-data for the previous calendar year (e.g. 2023) is available during October/November (2024), i.e., if we prepare for a new population update at the end of the year (2024), we will probably be able to get LPR-data for the full previous calendar year (2023), when we receive data in the beginning of the next year (2025). Other registries are updated at different time points (e.g., LMDB is updated quarterly).

Arendt JFH, Hansen AT, Ladefoged SA, Sørensen HT, Pedersen L, Adelborg K. Existing data sources in clinical epidemiology: Laboratory information system databases in denmark. Clin Epidemiol. 2020;12:469–75.

Christensen DH, Knudsen ST, Gylfadottir SS, Christensen LB, Nielsen JS, Beck-Nielsen H, et al. Metabolic factors, lifestyle habits, and possible polyneuropathy in early type 2 diabetes: A nationwide study of 5,249 patients in the danish centre for strategic research in type 2 diabetes (DD2) cohort. Diabetes Care. 2020;43(6):1266–75.

Christensen DH, Nicolaisen SK, Berencsi K, Beck-Nielsen H, Rungby J, Friborg S, et al. Danish centre for strategic research in type 2 diabetes (DD2) project cohort of newly diagnosed patients with type 2 diabetes: A cohort profile. BMJ Open. 2018;8(4):e017273.

Christensen H, Nielsen JS, Sørensen KM, Melbye M, Brandslund I. New national biobank of the danish center for strategic research on type 2 diabetes (DD2). Clin Epidemiol. 4:37–42.

Ehrenstein V, Antonsen S, Pedersen L. Existing data sources for clinical epidemiology: Aarhus university prescription database. Clin Epidemiol. 2010;2:273–9.

Gedebjerg A, Bjerre M, Kjaergaard AD, Nielsen JS, Rungby J, Brandslund I, et al. CRP, c-peptide, and risk of first-time cardiovascular events and mortality in early type 2 diabetes: A danish cohort study. Diabetes Care. 2023;46(5):1037–45.

Gedebjerg A, Bjerre M, Kjaergaard AD, Steffensen R, Nielsen JS, Rungby J, et al. Mannose-binding lectin and risk of cardiovascular events and mortality in type 2 diabetes: A danish cohort study. Diabetes Care. 2020;43(9):2190–8.

Grann AF, Erichsen R, Nielsen AG, Froslev T, Thomsen RW. Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at aarhus university, denmark. Clin Epidemiol. 2011;3:133–8.

Gylfadottir SS, Christensen DH, Nicolaisen SK, Andersen H, Callaghan BC, Itani M, et al. Diabetic polyneuropathy and pain, prevalence, and patient characteristics: A cross-sectional questionnaire study of 5,514 patients with recently diagnosed type 2 diabetes. Pain. 2020;161(3):574–83.

Hansen AL, Thomsen RW, Brøns C, Svane HML, Jensen RT, Andersen MK, et al. Birthweight is associated with clinical characteristics in people with recently diagnosed type 2 diabetes. Diabetologia. 2023;66(9):1680–92.

Johannesdottir SA, Horvath-Puho E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen HT. Existing data sources for clinical epidemiology: The danish national database of reimbursed prescriptions. Clin Epidemiol. 2012;4:303–13.

Jørgensen ME, Kristensen JK, Reventlov Husted G, Cerqueira C, Rossing P. The danish adult diabetes registry. Clin Epidemiol. 2016;8:429–34.

Kildemoes HW, Sørensen HT, Hallas J. The danish national prescription registry. Scand J Public Health. 2011;39(7 Suppl):38–41.

Kristensen FPB, Nicolaisen SK, Nielsen JS, Christensen DH, Højlund K, Beck-Nielsen H, et al. The danish centre for strategic research in type 2 diabetes (DD2) project cohort and biobank from 2010 through 2023 - a cohort profile update. 2024;16:641–56.

Mor A, Svensson E, Rungby J, Ulrichsen SP, Berencsi K, Nielsen JS, et al. Modifiable clinical and lifestyle factors are associated with elevated alanine aminotransferase levels in newly diagnosed type 2 diabetes patients: Results from the nationwide DD2 study. Diabetes Metab Res Rev. 2014;30(8):707–15.

Nielsen JS, Thomsen RW, Steffensen C, Christiansen JS. The danish centre for strategic research in type 2 diabetes (DD2) study: Implementation of a nationwide patient enrollment system. Clin Epidemiol. 2012;4(Suppl 1):27–36.

Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, Sørensen HT, Hallas J, Schmidt M. Data resource profile: The danish national prescription registry. Int J Epidemiol. 2017;46(3):798–798f.

Schmidt M, Pedersen L, Sørensen HT. The danish civil registration system as a tool in epidemiology. Eur J Epidemiol. 2014;29(8):541–9.

Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The danish national patient registry: A review of content,

data quality, and research potential. Clin Epidemiol. 2015;7:449–90.