

This document provides detailed information on the steps needed to replicate the full analysis in the paper **(Breaking) intergenerational transmission of mental health** by Aline Bütikofer, Rita Ginja, Krzysztof Karbownik and Fanny Landaud.

## Data access

The data employed in the analysis are drawn from Norwegian administrative registers. Researchers can gain access to the data by submitting a written application to Statistics Norway. The application should include a detailed research proposal describing the goals and methods of the project, a detailed list of variables, the selection criteria to be used, and how the research will be funded. Once received, applications must be certified in order to ensure that data are processed in a manner that protects the personal integrity of individuals surveyed. Conditional on this approval, Statistics Norway will then determine which data one may obtain in accordance with the research plan; all data are processed by Statistics Norway. Individuals must provide a list of all individuals who will have access to the data along with a description of a secure way in which the data will be stored for the period of the research project. Inquiries about access should be addressed to: Statistics Norway Postboks 2633 St. Hanshaugen, NO-0131 Oslo.

Researchers wishing to replicate our work should reference **project number 607173** "*Demografiske og sosiale endringer – utfordringer for sysselsetting og velferd*" at SSB. This project refers to the extracts made available to the Department of Economics - University of Bergen. This should be done after approval of the data from owners and the Regional Committees for Medical and Health Research Ethics. The sample includes everyone who resided in Norway in the period 1992-2015 and their parents (approx. 6.5 million people).

## Steps for replication

All the data preparation and analysis for this paper were performed using `Stata` version 16.1. The full replication package and raw data are stored at the servers of the University of Bergen. We provide the programs that allow to replicate the results of our paper.

## Required Stata packages

The full analysis was performed using `Stata` version 16.1.

## Memory and runtime requirements

The code was run on a 20-core (2 processors) Intel Xeon-based server with 1600 GB of RAM, Windows Server 2016 Standard.

## Directory structure

The replication package contains the following directories:

1. Folder named **final\_files** contains the programs used to create the data and the do files used in the analyses.
2. Folder named **tables\_rep** contains the output tables, graphs and log files with all the analyses results.

## Do-file structure

### Data preparation

The data preparation is easiest to replicate by running `master.do`. Before doing so, the researcher performing replication will have to change the path for the main directory to save the data and outputs of replication in each program where appropriate. All other paths written into `master.do` and the remaining do files should remain unchanged, as they will point to the file structure included in the replication package. The `master.do` runs the following do-files in the order listed (see also Table 1 below):

1. `clean_health.do`. Uses information from the *Control and Payment of Health Refunds* ("Kontroll og utbetaling av helserefusjoner", 2006-2020, Helsedirektoratet (2017a)) to construct the health indicators. These data are available from the Norwegian Health Directorate; see <https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/kuhr>. In order to access these datasets, approvals from the Regional Committees for Medical and Health Research Ethics are also required besides approval from the primary data owners.
2. `clean_npr.do`. Uses information from the *National Patient Register* ("Norsk pasientregister", 2008-2020, Helsedirektoratet (2017b)) to construct the health indicators. These data are available from the Norwegian Health Directorate; see <https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/norsk-pasientregister-npr>. In order to access these datasets, approvals from the Regional Committees for Medical and Health Research Ethics are also required besides approval from the primary data owners.
3. `clean_health_kids.do`. Uses the information from the output of `clean_health.do` to create specific health conditions and reduce sample size.
4. `middleschools`. Identifies the middle school where students completed 10th grade (last year of compulsory education) using information from the "Karakterer for avsluttet grunnskole".
5. `gpids`. Uses the information from the output of `clean_health.do` to identify the GP that each individual consulted most often each year.
6. `sample_pre`. Imposes the sample restrictions, and gathers information on sick leave for all individuals, demographic information, education, income and Social Security Records. Information about the data containing demographic characteristics (Statistisk Sentralbyrå, 2017a) can be found at <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/befolkning>. Highest completed education for each individual is available from the register for the Demographic information and highest education ("Demografiske opplysninger og høyeste utdanning") (Statistisk Sentralbyrå, 2017c). Data on individuals' income is collected by the Tax Authority (Statistisk Sentralbyrå, 2017b) and kept by Statistics Norway and documentation about these data can be found at Statistics Norway website <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/inntekt>. The information on employment is obtained from several register-based employment statistics, and in particular we rely on information from the Employer and Employee Register from the Norwegian Labour and Welfare Administration that is maintained by Statistics Norway (Statistisk Sentralbyrå, 2017d,e), and the documentation about these data can be found at <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/arbeidsmarked>.
7. The four programs in folder `with_sample_restrictions` use individual identifiers to construct the family network.

8. The following files merge health and education information for each relative in each family branch, identifying for each individual the parents, uncles/aunts and their spouses and the siblings of spouses of parents' siblings (and by age group in the parents' generation):

- `close.do, mor_cousin.do, mor_cousin_sp.do, far_cousin.do, far_cousin_sp.do, mor_sib_sp_sib.do, far_sib_sp_sib.do, close3035.do, mor_cousin_3035.do, mor_cousin_sp_3035.do, far_cousin_3035.do, far_cousin_sp_3035.do, mor_sib_sp_sib_3035.do, far_sib_sp_sib_3035.do.`

9. The following files gather information on health, education and occupation of parents, uncles/aunts, their spouses, and the siblings of spouses of parents' siblings and merge all these data to the focal child. The files are constructed for different ages in the parents' generation (25-30 and 30-35) and for different age ranges of the focal child (13-18 or 6-18). It also gathers information on grandparents' longevity using information from the mortality records ("Dodsarsaksregisteret" – DAR, Folkehelseinstituttet (2017). The documentation about these data can be found at <https://www.fhi.no/hn/helseregistre-og-registre/dodsarsaksregisteret/>.

- `merge25_30.do, merge30_35.do, merge25_30_age6.do, merge30_35_age6.do`

## Analysis

The second part of the file `master.do` includes the list of programs to run to produce all tables and figures in the paper. We describe next to each file the output that is produced after running it.

TABLE 1.  
Data Sources

| Data source  | Main description   | Variables   | Years available |
|--|--|---|-----------------|
| Control and Payment of Health Reimbursement (KUHR) | Individual primary care visits (to GPs or emergency rooms)             | Dates (year) of visits, with related diagnoses and symptoms (ICPC2) | 2006-2020       |
| Norwegian Patient Registry (NPR)                   | Individual inpatient and outpatient visits in specialist care.         | Entry & discharge dates (year), diagnoses (ICD10)                   | 2008-2020       |
| Mortality Records (DAR)                            | Individual death event   | Date (year) & cause of death (ICD10)                                | up to 2020      |
| Education Records                                  | For Parents  | Highest Completed Degree  | 1970-2020       |
| Education Records                                  | For Children   | Middle School GPA, Middle Schools Attended                          | 2020-2020       |
| Tax Authority Records                              | Annual information for eligible individuals (those formally employed). | Labor earnings and income from other sources                        | 1992-2020       |
| Population Records                                 | Annual demographic information about all individuals                   | Marital status, municipality of residence, gender, age, nationality | 1967-2020       |

*Note:* This table presents the time frame covered by the administrative records used in the analysis, the variables and years of data used.

## References

- Folkehelseinstituttet (2017). 'Dødsårsaksregisteret', <https://www.fhi.no/hn/helseregistre-og-registre/dodsarsaksregisteret/>, (last accessed: 11 July 2022).
- Helsedirektoratet (2017a). 'Kontroll og utbetaling av helserefusjoner-databasen, 2006-2016', <https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/kuhr>, (last accessed: 11 July 2022).
- Helsedirektoratet (2017b). 'Norsk pasientregister, 2008-2014', <https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/norsk-pasientregister-npr>, (last accessed: 11 July 2022).
- Statistisk Sentralbyrå (2017a). 'Befolkning Registret', <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/befolkning>, (last accessed: 11 July 2022).
- Statistisk Sentralbyrå (2017b). 'Inntekts- og formuesstatistikk for husholdninger, 1993-2014', <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/inntekt>, (last accessed: 11 July 2022).
- Statistisk Sentralbyrå (2017c). 'Nasjonal utdanningsdatabase, 1970-2014', <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/utdanning>, (last accessed: 11 July 2022).
- Statistisk Sentralbyrå (2017d). 'Registerbaserte sysselsettingsstatistikken', <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/arbeidsmarked>, (last accessed: 11 July 2022).
- Statistisk Sentralbyrå (2017e). 'System for persondata', <https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/arbeidsmarked>, (last accessed: 11 July 2022).