READ ME file for replication package for "The Physician-Patient Match and Health Inequality"

Ida Lykke Kristiansen and Sophie Yanying Sheng

September 23, 2025

This document provides information on how to replicate the analysis of the paper *The Physician-Patient Match and Health Inequality*. Because the data are confidential, we cannot share our data files. Instead, we provide information on how to access the data and provide all program files that generate the analysis data and the results of the paper.

The replication package consists of 45 Stata do-files and one SAS program file. Each file is described below.

1 Data

1.1 Data access

The results presented in the paper rely on a combination of several Danish administrative registers, which are merged using unique anonymous identifiers.

The use of these data is subject to the European Union's General Data Protection Regulation (GDPR). The data are stored on secure servers at Statistics Denmark and are not allowed to be transferred to computers outside Statistics Denmark. Researchers seeking access to the register data used in this paper must submit a written application for approval by Statistics Denmark. In the case of the prescription drug register and for the General Practice Providers register, approval from the Danish Health Data Authority is also required. Access may be granted to researchers affiliated with an accredited Danish research institution, or to international researchers collaborating with accredited Danish institutions. Once approval has been obtained, the data can be accessed remotely through Statistics Denmark's secure server. The complete list of databases required and who to contact to get access is provided in Table 1.

For more information see: https://www.dst.dk/en/TilSalg/Forskningsservice (accessed on September 18, 2025)

In addition to the Danish administrative registers, we also use survey data developed and collected by Mette Gørtz and Miriam Gensowski. We are grateful for access to their collected data and refer researchers to contact Mette Gørtz and Miriam Gensowski for questions regarding access.

2 Software

All data preparation and analysis were performed in either SAS (version 9.4 for Windows) or Stata (version 18.0 for Windows (64-bit)).

3 Replication files

The files reproduce all tables and figures reported in the paper and appendices. The first step of the work is done in SAS, while the rest is conducted in Stata. The Stata do-file `0_all.do` runs all Stata do-files in the right order and provides a short description of each file. In addition to the files described below the do-file `z_muni_codes.do` translates municipality codes across the 2007 municipality reform. The do-file `First` sets global paths.

Table 3 provides an overview of the figures and tables in the main manuscript, along with the do-files that generate them.

Table 4 provides an overview of the figures and tables in the online appendix, along with the do-files that generate them.

3.1 Preparing the raw data files: data_build.sas

This SAS program prepares all raw input data for the analysis. It extracts relevant subsets from Danish administrative registers, processes them, and exports .dta files for use in the Stata do-file workflow. To run the file, the user has to insert the relevant paths in the code (lines 1–4). The file then performs the following 13 steps:

1. Exports provider data.
2. Constructs patient cohorts: extracts children, links parents, and identifies education outcomes.
3. Creates parental education datasets.
4. Extracts mortality data.
5. Constructs the first steps in the patient–PCP matches using claims data.
6. Constructs PCP visits/services datasets from claims, including counts of visits, services, and diabetes check-ups.
7. Exports PCP costs from claims data.
8. Creates patient background data from the population register.
9. Extracts hospital admissions and diagnoses.
10. Extracts medication use.
11. Identifies lung cancer scans.
12. Exports income data and data on labor market participation.
13. Exports PCP high school grades.

Table 2 shows an overview of the input files, a short description of each step, and where the constructed dataset is later used. Table 1 provides an overview of the registers used and the identifiers applied for merging. *pnr* is the personal identifier; *ydernr* is the PCP identifier; *recnum* is hospital admission identifier.

3.2 Patient Background Information

After exporting the relevant data using the SAS file, we construct our patient data. This is done by running the following eight do-files in Stata. The first seven files each create datasets with specific patient characteristics (1a–1j), while the final do-file `1_patient_merge.do` merges the constructed datasets into one patient dataset.

Run the following do-files in the reported order to create patient-level background variables:

`1a_children_udd.do` – Takes the full population, identifies the highest completed education, and creates a dataset with information on gender, age, mother ID, and father ID.

`1b_mortality.do` – Identifies cause of death if the individual has died.

`1c_costs.do` – Extracts PCP reimbursements.

`1e_medicine.do` – Extracts use of prescription drugs.

`1f_hospital.do` – Extracts hospital admissions.

`1g_GP_services.do` – Extracts PCP services, visits, number of visits, mean services per visit, and diabetes check-ups.

`1j_lungscans.do` – Identifies lung scans from hospital data.

`1_patient_merge.do` – Cleans and merges information about the full population, including death, hospital admissions, and prescription drug use. Combines all datasets into one file, `patient_background.dta`.

3.3 Doctor background information

The do-files 2a–2e create datasets with specific doctor characteristics, which are merged in `2_GP_merge.do`. The do-files 2e–2h create additional datasets used in the doctor–patient match (`3a_doctor_patient_match.do`) or in analyses that are not part of the main analysis. These are explained below.

The intermediate datasets often include information for the full population, but this is later merged with the PCP data.

`2a_doctor_education.do` – Creates a dataset with all doctors in Denmark.

`2b_parent_edu.do` – Identifies all mothers and fathers in the full population, merges this with information on education and age, and saves the results in separate datasets. Later, these are merged with the PCP sample.

`2c_ydernrpnr.do` – Cleans data for PCP clinics. The raw dataset includes information on all people working in private clinics; we retain only doctors in primary care clinics. The final dataset contains personal identifiers, *pnr*, for each doctor working in each primary care clinic. The PCPs' personal identifier is then used to merge additional information about each PCP.

`2d_gp_grades.do` – Extracts high school grades.

`2e_gp_ethnicity.do` – Identifies ethnic background.

`2_GP_merge.do` – Merges PCP information and constructs two datasets for later use: one at the physician level (`Yder_9518_long.dta`) and one at the clinic level, where characteristics such as social background are aggregated at the clinic × year level (`gp_ses.dta`).

`2e_closures.do` – Identifies clinic closures based on the last date of services. The clinic closure information is used to identify doctor × patient matches (`3a_doctor_patient_match.do`) and as a source of variation in our main analysis.

`2g_parental_illness.do` – Identifies cases of parental illness among doctors. This dataset is used in a robustness check and merged later (`7d_parental_illness.do`).

`2h_parental_income.do` – Correlates parental education with parental income and parental unemployment. Correlations are reported in Section "2.1.1 Constructing the Analysis Sample," under the paragraph "Socio-Economic Status."

3.4 Merge Patients and Doctors

After constructing the physician and patient datasets, we merge the two into our main analysis sample and the full population sample used for summary statistics.

`3a_doctor_patient_match.do` – Implements the Kjaergaard algorithm to match patients and doctors. Note that the initial steps of the algorithm are performed in SAS during the initial data extraction process.

`3_merge.do` – Merges the patient and doctor data. Creates both a full population dataset and the main analysis dataset.

`3c_yder_pt_ses.do` – Calculates the share of low-SES patients per PCP–year. This variable is later used in `7c_PCP_Characteristics.do`.

`3d_PostClosureSESMeans.do` – Calculates the share of patients going to a low-SES PCP after a closure. Used in `8_pre_mortality_cliniclevel.do`.

3.5 Summary Statistics

These files construct the summary statistics reported either in the main manuscript or in the online appendix.

`4a_pt_sumstat.do` – Produces patient-level summary statistics (Table 1, Panel A; Table D1).

`4b_gp_sumstat.do` – Produces PCP- and clinic-level summary statistics (Table 1, Panels B–C).

`4c_test_for_selection.do` – Tests for PCP–patient selection (Table 2; Table D2).

`4d_test_for_selection_conditions.do` – Additional tests for PCP–patient selection (Table D3).

`4e_fourlineplot.do` – Plots the raw correlation between mortality and physician × patient SES (Figure D2).

`4f_sumstat_fig_gradient_mortality.do` – Plots the mortality gradient by patient educational level and PCP social background (Figure 1).

`4g_sumstat_fig_gradients.do` – Plots health–SES gradients (Figure D1).

3.6 Main Analysis

We then run the main analysis using the following three do-files:

`5a_main_results.do` – Main DDD regressions (Tables 3, A9, D4, D5).

`5b_event_figures.do` – Event-study figures (Figures 2, 3, A1).

`5c_detection_vs_adherence.do` – Health care utilization related to chronic conditions and detection vs. adherence (Tables D6–D8).

3.7 Robustness Checks

We test the robustness of the main results in the following files:

`6a_robustness_age.do` – Expanded age range (Table A1).

`6b_robustness.do` – Additional robustness tests (Tables A2–A4, A8).

`6c_robustness_SES_missing_incl.do` – Tests for missing values in the SES definition (Table A5).

`6d_soldclinic.do` – Uses clinic sales as a source of variation (Tables A6–A7).

3.8 Mechanisms

`7a_fragile_patients.do` – Tests the fragile-patients hypothesis (Table C1).

`7b_general_concordance.do` – Provides an external validation test (Table C2).

`7c_PCP_Characteristics.do` – Tests whether other PCP characteristics explain the results (Tables D10–D11).

`7d_parental_illness.do` – Tests the parental-illness hypothesis (Table D9).

3.9 Pre-Trends in Mortality (Appendix B)

`8_pre_mortality_cliniclevel.do` – Clinic-level tests (Figure B1, Panels C–D).

`8_pre_mortality_idlevel.do` – Individual-level tests (Figure B1, Panels A–B).

3.10 Survey data

The analysis using survey data was run on a different server, since the survey data was only available there, and the main part of the paper was not included on that server. The results were generated by using the do-files from this project to create the population data with the relevant variables, and then merging it with the survey data provided, which had already been cleaned by the data providers. The main results of the paper were recreated on the server with the survey data, confirming the results can be replicated on the other server.

`9_survey.do` - Physician-patient relationship using survey data (Table 4).

Table 1: Data sources and contact

Name	Database	Years	Identifier	Contact
BEF	Befolkningen (Population Demographics)	1995-2019	pnr	Statistics Denmark
DODSAASG & DODSAARS	Dødsårssagsregistret (Cause of death registry)	1970-2019	pnr	Statistics Denmark
FAIN	Husstande og familier (Households and families)	1995-2007	pnr	Statistics Denmark
IDAP	Arbejdstilknytning (The Integrated Database for Labour Market Research)	1980-2019	pnr	Statistics Danmark
LMDB	Lægemiddeldatabasen (The Danish National Prescription Registry)	1995-2019	pnr	The Danish Health Data Authority
LPR	Landspatientregistret (The National Patient Registry)	1995-2019	pnr & recnum	Statistics Denmark
SSSY & SYSI & SYSI	Sygesikring (The National Health Insurance Service registry)	1995-2019	pnr & ydernr	Statistics Denmark
UDDA	Uddannelser (Educations)	1980-2019	pnr	Statistics Denmark
UDG	Karakterer for afsluttede uddannelser (Grades on completed educations)	1985-2019	pnr	Statistics Denmark
YDER	Yderregister (General Practice Providers)	1995-2019	pnr & ydernr	The Danish Health Data Authority

Table 2: Overview on SAS steps and input files and later use in Stata

Step	Description	SAS Input	Used Later in Stata
1. Provider data	Exports provider registers	ext.T_yder_1	1a_doctor_education.do
		ext.T_yder_person	2a_doctor_education.do
		ext.Registration	2c_ydernrpnr.do
2. Patient cohorts	Creates dataset on children and educational outcomes	raw.Fainupdv	1a_children_udd.do
		raw.uddaupd02v	2a_doctor_education.do
3. parental education	Finds highest completed education of parents	raw.uddaupd02v	2b_parent_edu.do
4. Mortality	Find cause of death	raw.Dodsaasgv	1b_mortality.do
		raw.Dodsaarsv	
5. Patient–PCP match	Constructs first steps in the patient–PCP matches	raw.Sssyupd02v	3a_doctor_patient_match.do
		raw.Sysiupdv	
6. PCP services	Finds PCP visits, services, diabetes checks	raw.Sssyupd02v	1g_GP_services.do
		raw.Sysiupdv	
7. PCP costs	Exports PCP costs from claims data.	raw.Sssyupd02v	1c_costs.do
		raw.Sysiupdv	
8. Patient background	Creates patient background data	raw.Befupd01v	1_patient_merge.do
			2f_ietype.do
9. Hospital admissions	Finds hospital admissions & diagnoses	raw.lpradmv	1f_hospital.do
		raw.lprdiagupdv	
10. Prescriptions	Finds prescription medicine use	lmdb.lmdbYYYY (1995–2021)	1e_medicine.do
			2g_parental_illness.do
11. Lung cancer scans	Finds operation codes for lung scans	raw.Lprsksubupdv	2j_lungscans.do
		raw.lpradmv	
12. Income	Finds parental income	raw.Idapupdv	1h_parental_income.do
		raw.Indupd02v	
13. PCP grades	Finds high school grades	raw.Udgv	1d_gp_grades.do

Table 3: List of output from the main paper and the do-files creating the output

Output	Title	do-file
Figure 1	One-Year Mortality by Patient Education and Physician SES	`4f_sumstat_fig_gradient_mortality.do`
Figure 2	The Effect of Physician-Patient SES Concordance on Mortality	`5b_event_figures.do`
Figure 3	The Effect of Physician-Patient SES Concordance on Health Behaviors	`5b_event_figures.do`
Table 1	Summary Statistics - Patients, Physicians, and Clinics	`4a_pt_sumstat.do` `4b_gp_sumstat.do`
Table 2	Test for Selection in Physician-Patient Reassignment After Clinic Closure	`4c_test_for_selection.do`
Table 3	The Effect of Physician-Patient SES Concordance on All-Cause Mortality and by Cause	`5a_main_results.do`
Table 4	The Effect of Physician-Patient SES Concordance on Physician-Patient Relationship	`9_survey.do`

Table 4: List of output from the Online appendix and the do-files creating the output

Output	Title	do-file
Figure A1	The Effect of Physician-Patient SES Concordance on Number of Visits in Different Samples	5b_event_figures.do
Table A1	Robustness Check: Expanding age range to 30-80	6a_robustness_age.do
Table A2	Robustness Check: The Effect of SES Concordance When Excluding Non-ethnic Danish Patients	6b_robustness.do
Table A3	Robustness Check: Alternative Physician SES Aggregation to the Clinic Level	6b_robustness.do
Table A4	Robustness Check: Using a Subsample of Physicians with Non-missing SES	6b_robustness.do
Table A5	Robustness Check: Including physician missing SES information separately	6c_robustness_SES_missing_incl.d
Table A6	Robustness Check: Clinics Acquired by a New Physician - Test for selection	6d_soldclinic.do
Table A7	Robustness Check: Clinics Acquired by a New Physician	6d_soldclinic.do
Table A8	Robustness Check: Conditional on Survival and No Subsequent Physician Switching	6b_robustness.do
Table A9	The Effect of Physician-Patient SES Concordance on Mortality	5a_main_results.do
Figure B1	Placebo test - Predicted SES PCP at the individual level (id) or clinic level on pre-closure mortality	8_pre_mortality_cliniclevel.do 8_pre_mortality_idlevel.do
Table C1	Mechanism: The Effect of Having a Low-SES Physician (PCP) on Mortality by Patients with Different Baseline Conditions	7a_fragile_patients.do
Table C2	External Validity: The Effect of Physician Parental Educational Level on Patient Mortality by Patient Education	7b_general_concordance.do
Figure D1	Health-SES Gradient by Outcomes of Interest	4g_sumstat_fig_gradients.do
Figure D2	Physician (PCP)-Patient SES Concordance and Mortality - Raw Correlations	4e_fourlineplot.do
Table D1	Summary Statistics - Patients	4a_pt_sumstat.do
Table D2	Duration of Patient-Physician Relationships Following Clinic Closure	4c_test_for_selection.do
Table D3	Test for Selection in Patient-Physician Reassignment	4d_test_for_selection_conditions
Table D4	The Effect of Physician (PCP)-Patient SES Concordance on Mortality from Chronic Conditions by Gender and Age	5a_main_results.do
Table D5	The Effect of Physician-Patient SES Concordance on Healthcare Utilization	5a_main_results.do
Table D6	The Effect of Physician (PCP)-Patient SES Concordance on medicine use	5c_detection_vs_adherence.do
Table D7	The Effect of Physician-Patient SES Concordance on Health Care Utilization, Disease Detection, and Treatment Adherence	5c_detection_vs_adherence.do
Table D8	The Effect of Physician-Patient SES Concordance on Mortality	5c_detection_vs_adherence.do
Table D9	Mechanism: The Effect of Physicians' (PCP) Parents' Illness on Patient Mortality	7d_parental_illness.do
Table D10	Internal Validity: The Role of Other Physician (PCP) Characteristics in Reducing the SES-Gradient in Mortality	7c_PCP_Characteristics.do
Table D11	Internal Validity: The Effect of Physician (PCP)-patient SES Concordance on Mortality by Physician Characteristic	7c_PCP_Characteristics.do