**Code Documentation**

Need to do:

-Document the drug to ATC4 mapping

-Drug -> ATC -> brand name/Part D.

**Raw Data Folders and File from NBER**

* */disk/aging/partd/cutler-DUA28717/20pct/pde/`year’/pde`year’.dta*: Part D Event Files. When a Medicare beneficiary with Part D coverage fills a prescription, the prescription drug plan submits a record to CMS. The PDE file includes all transactions covered by the Medicare prescription drug plan for both Prescription Drug Plans (PDPs) and Medicare Advantage Prescription Drug Plans (MA-PDs). The fields GNN (generic name) PROD\_SRVC\_ID (NDC) allow linkage with the ATC4 categories from *ndc\_map.* List of included variables here: <https://resdac.org/cms-data/files/pde/data-documentation> (including Patient Gender, DoB, CCW/NCPDP Pharmacy Identifier, Prescriber Identification Number e.g. NPI). Available for years <=2017.

**Raw Data Folders and Files**

* *data/DAC\_NationalDownloadableFiles/DAC\_`mm’\_`yyyy’.csv*: These datasets give demographic information about doctors, clinicians and groups listed on Medicare Care Compare, from 2014 to 2020. Each observation is unique at the clinician/enrollment record/group/address level. Variables include NPI, organization legal name, organization PAC ID, number of group members, as well as some credentials of interest like the clinician’s medical school and graduation year. Taken from <https://data.cms.gov/provider-data/dataset/mj5m-pzi6>. Note that archived data exist for four months for each available year. I merge all these month-year data sets together when creating the “master” data set to obtain the most comprehensive union of enrolled physicians.
* *data/PartD/ProviderandDrug/Provider`year’.csv*: The Medicare Part D Prescribers by Provider dataset contains information on prescription drugs prescribed by individual physicians and other health care providers and paid for under the Medicare Part D Prescription Drug Program. The dataset identifies providers by their National Provider Identifier (NPI) and summarizes for each prescriber the total number of prescriptions that were dispensed, which include original prescriptions and any refills, and the total drug cost. Includes data on the following drug categories: opioids and antipsychotics. Taken from <https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider>, which is at the provider-year-level.
* *data/PartD/ProviderandDrug/ProviderDrug`year’.csv* and *data/PartD/ProviderandDrug/PartDProviderDrug`year’.dta*: The primary data source is CMS Chronic Condition Data Warehouse, a database with 100% of Medicare enrollment and final-action Part D prescription drug event (PDE) data. The study population includes utilized drugs from Medicare Part D PDEs submitted by the Part D sponsors for the benefit year. The data are aggregated based on all Medicare Part D PDE data, excluding over the counter prescriptions. Taken from <https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider-and-drug>, which is at the provider-drug-year-level.
* *data/PartB/ProviderandService/ProviderSrvc`year’.csv*: The primary data source is CMS Chronic Condition Data Warehouse, a database with 100% of Medicare enrollment and fee-for-service claims data. Service counts, beneficiary counts, provider charges, Medicare allowed amounts and payments and the place of service indicator are summarized from Part B non-institutional claims processed through Medicare Administrative Contractor Jurisdictions. The study population includes providers that had a valid NPI and submitted Medicare Part B non-institutional claims (excluding DMEPOS) during the reporting period. Taken from <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service>, which is at the provider-service-year level.
* *data/PartD/OpioidRates/opioid\_rates\_region.csv*: The number and percentage of opioid prescriptions filled in the Medicare Part D program, by region (national/state/county/zip-code) from 2013-2019. <https://data.cms.gov/summary-statistics-on-use-and-payments/medicare-medicaid-opioid-prescribing-rates/medicare-part-d-opioid-prescribing-rates-by-geography/data>
* *data/PartB/Provider/Provider`year’.csv*: The Medicare Physicians & Other Practitioners by Provider dataset provides information on use, payments, submitted charges and beneficiary demographic and health characteristics organized by NPI. This dataset is based on information gathered by the CMS administrative claims data for Original Medicare Part B beneficiaries available from the CMS Chronic Conditions Data Warehouse. Taken from <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider>.
* *data/PartB/Service/Service`year’.csv*: Information on services and procedures (HCPCS codes) provided to FFS Medicare Part B beneficiaries by physicians and other providers; aggregated by geography (national or by state) and service. Taken from <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-geography-and-service>. Includes average Medicare payment amount.
* *data/ZipCrosswalks/ziphsahrr20132019.dta*: Allows me to aggregate data from the ZIP code level to the hospital service area (HRR) or hospital referral region (HRR) level. Taken from <https://data.dartmouthatlas.org/supplemental/>.
* *data/HealthIT/heathit\_attest.csv*: The Medicare Electronic Health Record (EHR) Incentive Program provides incentives to eligible clinicians and hospitals to adopt EHR. This data set identifies the unique vendors/products/product types of each certified health IT product used by participating providers to attest to meaningful use. It allows us to see which NPI has attested to meaningful IT use. Years 2011-2017. Taken from <https://www.healthit.gov/data/datasets/ehr-products-used-meaningful-use-attestation>.
* *data/QPP/QPP\_Experience\_Report\_`y’.csv*: De-identified MIPS data from 2018-2020, with data on payment adjustments, total allowed charges, and components being broken down into more detailed parts. <https://data.cms.gov/quality-of-care/quality-payment-program-experience>
* *data/DAC\_NationalDownloadableFiles/patient\_exp2019.csv*: Contains the Consumer Assessment of Healthcare Providers and Systems (CAHPS) for Merit-based Incentive Payment System (MIPS) measures submitted by groups. Years 2019+2020 only. Data is at organization-CAHPS group level. Taken from <https://data.cms.gov/provider-data/dataset/8c70-d353>.
* *data/DAC\_NationalDownloadableFiles/mips\_perf`y’.csv*: Contains the MIPS final scores and performance category scores for clinicians participating. Data is at the NPI-organization level. For years 2019+2020. Taken from <https://data.cms.gov/provider-data/dataset/a174-a962>.
* *data/merged/PartDPartBDACMIPS\_flags.dta*: Contains the union of all NPIs that appear in Part D, Part B, DAC and MIPS data – as well as flags indicating which data set they are in.
* *data/MappingData*: Contains HRR and PCSA level shapefiles from the Dartmouth Atlas <https://data.dartmouthatlas.org/supplemental/>, <https://data.world/dartmouthatlas/pcsa-data-crosswalk>.

**Code Files**

* *ndc\_map.R*: Map the FDA NDCs to Drug Classes and Terminologies by querying the RxNorm API at <https://rxnav.nlm.nih.gov>.
* *join\_dac\_nationaldownloadablefiles.R*:Prepare the demographic data for physicians enrolled in Medicare Care Compare, i.e. in *data/DAC\_NationalDownloadableFiles*. The .R code “aggregates” the raw month-year data sets *DAC\_’mm’\_’yyyy’.csv* to the yearly level.
* *match\_drug\_to\_atc.do*: Match the drugs that appear in the Part D Provider-Drug files with the associated ATCs.
* *create\_docpracticelvl\_data.do*: Creates master data sets of provider/prescriber-level characteristics. Also computes the Part D/Part B Z-scores for analysis.
* *analyze\_provider\_data.do*: Conduct analyses using the data sets created by *create\_docpracticelvl\_data.do*

**Created Data Files**

* *data/PartD/ndc\_map*: All unique NDCs from NBER’s Part D Event files */disk/aging/partd/cutler-DUA28717/20pct/pde/`year’/pde`year’.dta*, plus a column for the associated ATC4 codes.
  + *ndc\_allatc4\_categories\_map.dta* pools together the *ndc\_atc\_map`year’* and links each NDC to all the associated ATCs and drug categories.
  + *ndc\_bn\_gnn\_from\_pdefiles.dta* contains all the NDCs that appear in the NBER PDE files as well as their associated generic names and brand names.
  + *ndc\_allatc4\_categories\_bnn\_gn\_map.dta* is just *ndc\_allatc4\_categories\_map.dta* plus the generic and brand names of drugs associated with the NDCs.
* *data/DAC\_NationalDownloadableFiles/DAC\_NatonalDownloadableFiles20142020.dta*: The datasets give demographic information about doctors, clinicians and groups listed on Medicare Care Compare, from 2014 to 2020. Each observation is unique at the clinician/enrollment record/group/address level. Variables include NPI, organization legal name, organization PAC ID, number of group members, as well as some credentials of interest like the clinician’s medical school and graduation year. Taken from <https://data.cms.gov/provider-data/dataset/mj5m-pzi6>. Note that archived data exist for four months within each available year. Using the script *join\_dac\_nationaldownloadablefiles.R* and then the program *join\_dac\_ndfs* in *create\_docpacticelvl\_data.do*, I merge all these month-years when creating the “master” data set *data/DAC\_NationalDownloadableFiles/DAC\_NatonalDownloadableFiles20142020.dta* in order to obtain the most comprehensive union of enrolled physicians.
* *data/DAC\_NationalDownloadableFiles/DAC\_NDF20142020\_grpchanges.dta*: Keeps track of group changes among physicians that appear in the DAC data. Created using *DAC\_NatonalDownloadableFiles20142020.dta* and the program *join\_dac\_ndfs* in *create\_docpacticelvl\_data.do*.
* *data/DAC\_NationalDownloadableFiles/DAC\_apm\_grp\_linkage.dta*: Link NPIs to their respective PACID groups as well as APM groups. The APM groupings are artificially constructed by the function *clean\_mips\_healthit* in *crea\_docpracticelvl\_data.do*.
* *data/DAC\_NationalDownloadableFiles/mips\_perf20192020\_`source’.csv*: Contains MIPS scores for individuals, groups and APMs, with imputed scores for PI. Also contains tags for whether the entity (NPI) receive just the minimum score for neutral payment (30 in 2019, 45 in 2020) or have full information for every component. Created by the function *clean\_mips\_healthit* in *create\_docpracticelvl\_data*.
* *data/PartD/ProviderandDrug/PartDProviderDrugCtgrs20132019.dta*: This merges the raw CMS Part D Provider-and-Drug files with the NDC-drug name-drug category mapping *ndc\_allatc4\_categories\_bnn\_gn\_map.dta*, using the program *import\_providerdrug\_partd* in *create\_docpracticelvl\_data.do*. It gives us how many prescriptions of each category an NPI prescribes in each observed year. **Note that the categories in this file do not contain opioids, as opioid total claims will be available in the Part D Provider Dataset.**
* *data/PartD/ProviderandDrug/PartDProvider\_AllDrugCtgrs20132019.dta*: **This is the main Part D data set we will work with and is at the NPI-group-year level.** To create this, we use the program *import\_provider\_partd* in *create\_docpracticelvl\_data.do* tomerge the CMS Part D Prescriber data set *data/PartD/ProviderandDrug/Provider`year’.csv* with the above data *data/PartD/ProviderandDrug/PartDProviderDrugCtgrs`year’.dta* on total claims of different drug categories, this time including opioids, antibiotics and antipsychotics. It also includes some basic prescriber demographic information. We then merge with the Medicare Compare data in *DAC\_NatonalDownloadableFiles20142020.dta* to link physicians to group practices in each year. Finally, we merge with supplementary date sets such as Health IT/MIPS scores and group/HRR-lvel characteristics.
* *data/PartD/ProviderandDrug/PartDProvider\_AllDrugCtgrs\_AvgClaimsZScores\_20132019.dta*: This is *data/PartD/ProviderandDrug/PartDProvider\_AllDrugCtgrs20132019.dta,* at the **NPI-grou-year** level plus prescribers’ pooled z-scores, calculated using the program *create\_partd\_zscores* in *create\_docpracticelvl\_data.do*..
* *data/PartB/ProviderandService/ProviderSrvc\_EMCodeCounts.dta*: This contains the HCPCS\_code service count variable from *data/PartB/ProviderandService/ProviderSrvc`year’.csv* billed by the NPIs in each observed year. It is created using the program *import\_providerservice\_partb* in *create\_docpracticelvl\_data.do*. It also separates out the HCPCS\_codes corresponding to E/M codes, as well as dividing them into finer sub-categories of E/M like office visits, hospital inpatient, etc.
* *data/PartB/ProviderandService/PartBProvider\_AllServiceCounts20132019.dta*: **This is the main Part B data set we will work with and is at the NPI-group-year level**. To create this, we use the program *import\_provider\_partb* in *create\_docpractice\_lvl\_data.do* to merge the CMS Part B Prescriber data set *data/PartB/Provider/Provider`year’.csv* (with prescriber-year-level demographic information, as well as total service counts, total beneficiaries receiving services, total medical/drug services) with the data set above *data/PartB/ProviderandService/ProviderSrvc\_EMCodeCounts20132019.dta* on E/M code counts. We then merge with the Medicare Compare data in *DAC\_NatonalDownloadableFiles20142020.dta* to link physicians to group practices in each year. Finally, we merge with supplementary date sets such as Health IT/MIPS scores and group/HRR-lvel characteristics.
* *data/PartB/ProviderandService/PartBProvider\_AllServiceCounts\_UpcodingZScores20132019.dta*: This is *data/PartB/ProviderandService/PartBProvider\_AllServiceCounts20132019.dta*, at the **NPI-group-year level** plus providers’ pooled z-scores, calculated using the program *create\_partb\_zscores* in *create\_docpracticelvl\_data.do*.
* *data/PartB/Service/StateLevel\_MediumSrvcPrices.dta*: Cleans up *data/PartB/Service/Service`year’.csv* to get a comprehensive dictionary of state-level prices for medium-intensity E&M codes from 2013 to 2019, using the program *get\_partb\_prices* in *create\_docpracticelvl\_data.do*.
* *data/PartB/EMRevenues.dta*: Contains ingredients for computing the “into money measure” for Part B providers, which include providers’ total revenues from E&M services as well as their counterfactual revenues if they had coded everything at the respective medium intensity level. These data are **at the provider-year level**. Created using the program *create\_partb\_intomoney\_measure* in *create\_docpracticelvl\_data.do*.
* *data/ZipCrossWalks/HRR\_characteristics.dta*: Comprises a set of relevant HRR-year-level features, such as the amount of home health activity going on. This file is created using the program *get\_hrr\_chars* in *create\_docpracticelvl\_data.do.*
* *data/ZipCrossWalks/PCSA\_characteristics.dta*: Comprises a set of relevant PCSA-year-level features, such as the amount of home health activity going on. This file is created using the program *get\_pcsa\_chars* in *create\_docpracticelvl\_data.do.*
* *data/DAC\_NationalDownloadableFiles/group\_characteristics.dta*: Comprises a set of relevant group-year-level features, such as group size, % graduates of top 20 med schools, patient experience and performance measures. This file is created using the program *get\_grppractice\_chars* in *create\_docpracticelvl\_data.do*.
* *data/HealthIT/healthit\_attest\_cleaned.dta*: Health IT data cleaned—includes all the NPIs in the original data set *data/HealthIT/heathit\_attest.csv*, and various indicators of EHR adoption.

**Guide: How to Put Drugs into Categories**

1. Get a list of all NDCs using the program *get\_ndc\_from\_pde* in *match\_drug\_to\_atc.do*, which gets a list of all unique NDCs from the yearly PDE files from NBER.
2. Run *ndc\_map.R* on the files created in (1) to make files mapping the NDCs to ATC4s.
3. Run the program *clean\_up\_ndc\_atc\_map* in ­*match\_drug\_to\_atc.do* to clean this mapping and create a “master map” by pooling all years together. That is, for each NDC, create a variable containing all the associated ATC4s as well as other variables indicating the relevant drug categories.
4. Get all the NDCs and the associated brand and generic names using the program *get\_bn\_gnn\_from\_pde* in *match\_drug\_to\_atc.do.*
5. Use the program *merge\_ndc\_drug\_names* in ­*match\_drug\_to\_atc.do*, which merges *ndc\_allatc4\_categories\_map.dta* with *ndc\_bn\_gnn\_from\_pdefiles.dta* to create *ndc\_allatc4\_categories\_bnn\_gn\_map.dta*. This maps each NDC to ATC4s, a drug category, a brand name, and a generic name.
6. Merge this mapping with the Part D Provider and Drug data set.

**Guide**: **How to Create the DAC National Downloadable Files**

1. Download the monthly archived data (there are four months with archived data for every year from 2014-2020).
2. Run *join\_dac\_nationaldownloadablefiles.R* to union, for each year, these four months and aggregate them up into the corresponding yearly file.
3. Run the program *join\_dac\_ndfs* in *create\_docpracticelvl\_data.do* to (a) restrict to specific variables of interest, and (b) append the different years together to create a master physician-level data set.

**How to Use NBER Remote Computer**

1. Log into FortiClient.

Username: lucasdo

Password: password associated with [lucasdo@nber.org](mailto:lucasdo@nber.org)

Meomeo53Meomeo11!

1. Log into Remote Desktop Connection.

Username: lucasdo-dua28717

Password: above.

1003001363