Quick Start Guide for Inovalon*

February 24, 2025

Click Here for the Most Recent Version

*__Disclaimer__: The content of this book is for informational purposes *only*, and policies regarding access to and use of Inovalon may change over time. For the most up-to-date and accurate information, we *strongly* recommend contacting the appropriate authority. Please note that an HBS email account is required to access some of the links included in this document.

Revision Log

1. Version 1.0 draft created by Jinyeong Son in October, 2024.
2. Version 2.0 draft created by Jinyeong Son in February, 2025.

Contents

1 Overview

Inovalon is a payer-sourced healthcare database that contains claims data from more than 160 payers or providers of healthcare benefits and administrative services. The key data scopes are as follows:

- **Payer Segment**:
  - Commercial (group/individual and self/fully-insured),
  - Medicaid Managed Care (*not* Medicaid FFS),
  - Medicare Advantage (*not* Medicare FFS and Medigap), and
  - Dual Eligibles.
- **Sample Period**: Medical and prescription drug claims paid between 2010 and 2020.
- **Geographic Coverage**: All 50 states, the District of Columbia, and Puerto Rico.

Inovalon provides information on 217.4 million individuals throughout the sample period across all payer segments. More specifically, the average *annual* number of covered lives is approximately 46.8 million, with this number increasing over time (e.g., from approximately 17.8 million in 2010 to 71.8 million in 2019).[1,2] Section 6.1 presents further details and additional statistics on covered lives by segment, year, and state.

Inovalon is a *panel* dataset that allows researchers to follow specific individuals over time across different segments using an anonymized identifier, provided they continue holding health insurance plans from any of the sourced payers.

2 Core Datasets and Key Variables

Inovalon comprises five core datasets: (1) member, (2) enrollment, (3) medical claims, (4) prescription drug claims, and (5) provider. Below, we provide a detailed description of each dataset, along with the key variables contained within them.

[1] We construct these numbers based on the baseline definition in Section 6.1—individuals with insurance as of *July 1st* in a given year—along with some restrictions. Specifically, we include only individuals enrolled in plans that offer both medical and pharmacy benefits. Additionally, we consider only enrollment periods during which both medical and pharmacy claims are available.

[2] In 2020, the number of covered lives was 53.9 million, deviating from the increasing trend due to significant disruptions in data collection caused by the pandemic.

2.1 Member

The member file contains demographic information for each individual and provides the following five variables:

- year of birth, gender, state, three-digit ZIP code, and race/ethnicity.

2.2 Enrollment

The enrollment file includes detailed information about health insurance plans each individual has enrolled in. Specifically, it provides each enrolled plan's coverage period, along with information on broader and more detailed plan segments:

- effective and termination dates, payer group (commercial, Medicaid, or Medicare), and payer type (e.g., Medicaid low income or Medicaid disabled among Medicaid enrollees).

In addition, it contains various attributes for enrolled insurance plans, including but not limited to:

- product type (e.g., HMO, PPO, or POS),
- information on group plans (i.e., individual, small group, and large group) for commercial plans, and
- ACA-related variables (e.g., an indicator for whether a plan is an ACA plan, or an indicator for whether a plan is offered on the ACA exchange).

2.3 Medical Claims

The medical claims file contains line-level data on medical services provided. Using this data, researchers can obtain the following basic information:

- provider information,
- payment status of claim (e.g., initial, denial, or adjustment),
- service beginning/end date,[3]

[3]For inpatient room and board claims, the service start and end dates can be used as the admission and discharge dates, respectively.

- financial information (e.g., allowed amount,[4] copay amount, or paid amount),[5] and
- type of services (i.e., inpatient, outpatient, and professional).

The medical claim code file, which exists separately, can be merged with the medical claims file to supply additional information, such as:

- DRG,
- HCPCS/CPT,
- ICD-9/ICD-10, and
- POS (place of services, e.g., emergency room).

It is important to understand that the medical claim code file is the largest dataset in Inovalon. This is because it is a long-form file in which each medical claim spans multiple rows, with each row containing unique information associated with the claim (e.g., CPT, ICD, POS, etc.). Therefore, we recommend that researchers specify the subsets (e.g., only CPT) they need to extract; otherwise, the dataset may become excessively large.

2.4 Prescription Drug Claims

The prescription drug claims file includes variables pertaining to pharmacy (or prescription) claims. Specifically, it contains the following information:

- provider information,
- payment status of claim (e.g., initial, denial, or adjustment),
- prescription fill date, 11-digit NDC, and supply days, and
- financial information (e.g., allowed amount, copay amount, or paid amount).

2.5 Provider

As indicated above, both the medical and prescription data files include provider information, specifically an anonymized provider identifier. By merging with the provider file, researchers can access additional details about the provider. The provider file contains the following variables:

[4] The amount the insurance company allows the provider to charge under contract with the provider for the service performed.

[5] The amount the insurance company actually paid to the provider for the service performed.

- provider, business, and parent organization names,
- NPI,
- practice and billing addresses, and
- taxonomy (i.e., type, classification, and specialization).

It is worth noting that the provider file is generated through a validation process by a third party. Researchers can also use the provider *supplemental* file, which includes similar provider details but comes directly from the underlying data sources without undergoing the validation process.

3 Some Important Features of Data

3.1 Complete Claims History

For both the medical and prescription drug claims files, Inovalon offers two distinct versions—a *non-consolidated* version and a *consolidated* version. The key distinction between the two versions is whether they include the full history of each claim (non-consolidated version) or retain only the final action on the claim (consolidated version).

Researchers have the flexibility to choose which version of the data to use, depending on their research questions. For instance, if one is interested in using claim denials as an outcome variable, they should opt for the non-consolidated version. Conversely, if only the final claim is of interest, researchers should use the consolidated version to remove duplicates when aggregating certain variables (e.g., costs or days supply).

3.2 Geographic Information

As briefly described in Section 2.1, the member file provides each individual's state of residence and three-digit ZIP code. Researchers should exercise caution when using this information, as it is recorded at a single point in time. In other words, the location variables are not tracked longitudinally, so researchers cannot link this information to each enrolled plan *separately*. Instead, they have only one location data point for all enrolled plans, which may not be correct if an individual moves.[6]

[6]For instance, I lived in Texas from 2018 to 2023 and moved to Massachusetts in 2024. Suppose that I had health insurance plans from two distinct firms, both of which are data contributors to Inovalon. I would appear as two rows in the enrollment file, as I had two different plans. However, the member file provides

One way to address this challenge is to infer an individual's location based on the provider's practice location using the individual's claims, a method commonly used in movers research design (e.g., Finkelstein, Gentzkow and Williams (2016)).

3.3 Financial Information

Inovalon *prohibits* researchers from combining provider information (e.g., NPI) with *actual* financial information (e.g., allowed amount or paid amount). Specifically, if a claim includes provider information, it excludes financial details, and vice versa.

Additionally, even when claims contain financial information, certain details (e.g., copay amount) may still be missing. Furthermore, when zero values are recorded in the financial data fields, it can be difficult to determine whether they represent actual zero values or missing data.

3.4 Linkage to External Data

Inovalon allows researchers to link their data with external datasets at the *member* (or *beneficiary*) level through Datavant, with which Inovalon has a partnership. Researchers may purchase additional member-level socioeconomic status information (e.g., income, assets, etc.) or other commonly used datasets for merging, both of which are already tokenized by Datavant. Otherwise, researchers must first obtain approval from the agency providing the data to have it go through the tokenization process by Datavant, after which it can be linked to Inovalon.

As stipulated in the Data Use Agreement (see Section 4.2 below), the authorities at HBS, Chirag Patel and Rachel Talentino, must be notified first, and the process must undergo a review to determine whether it breaches any clauses in the agreement with Inovalon or HIPAA regulations. Therefore, please contact them before proceeding with any linkage to external data.

In contrast to the member-level linkage, at the *provider* level, we obtained approval from Inovalon for researchers to merge *publicly* available provider data (e.g., the National Plan and Provider Enumeration System (NPPES) data) with Inovalon using NPI when necessary.

only a single data point for my location, either TX (if measured in 2023 or earlier) or MA (if measured in 2024).

4 Eligibility and Onboarding Process

4.1 Eligibility

Individuals affiliated with HBS, including both faculty and students, are eligible to request access to Inovalon. Students will need a PI-eligible faculty sponsor to conduct independent research using Inovalon, as the onboarding process detailed below requires submitting an IRB, which mandates designating a faculty member as the PI.

4.2 Onboarding Process

The onboarding process primarily consists of two parts: completing the IRB process and setting up an HBS Grid working directory. Potential users can proceed with the necessary steps for both parts *simultaneously*. Below, we provide further details on the required steps for each part separately.

IRB Process Researchers need to complete the following steps sequentially to finish the IRB process:

1. Review and Sign the Data Use Agreement (DUA),
2. Create a DUA record for the project and link it to the HBS Inovalon Master DUA record (DUA22-1285),
3. Obtain approval for the IRB and Data Safety.

Please note that, aside from signing the DUA, the remaining IRB process will be completed through Harvard ESTR, and the main contact for this process is Rachel Talentino. In addition, if the IRB review board requires specific training, researchers can complete it through the CITI Program. We also have an HBS SharePoint webpage dedicated to this process, where users can find more details about the procedure.

HBS Grid Set-Up Inovalon data are classified as Level 4, meaning researchers can access the requested data only through HBS Research Computing Services (also known as HBS Grid). To set up a working directory for the project, researchers must first create an HBS Grid account, followed by setting up a VPN through the installation and configuration of the required application. For more details, please visit the HBS Grid webpage.

4.3 Data Access Flow

The original copy of the Inovalon database is housed at Harvard Medical School (HMS) and managed through Microsoft SQL Server Management Studio. Researchers first consult with Chirag Patel to describe and determine the data needed for their research. He will then retrieve the data from HMS and place the extracts into the researcher's designated directory on HBS Grid. Technically, researchers do not need to write SQL code for data extraction; however, it is recommended that they be able to read it to verify that the extracted data are consistent with their request.

5 Some Important Points Researchers Should Know

5.1 Data Availability Period

Inovalon provides four *undocumented* variables—claim adjusted effective/termination dates and Rx claim adjusted effective/termination dates—which researchers must consider before conducting any analysis. This is because plan enrollment periods may differ from claims data availability periods, as the data-providing agency may stop submitting claims to Inovalon for various reasons.

To clarify this point, let us illustrate with a simple example. Suppose your sample period is from 01/01/2020 to 12/31/2020, and you identify an individual as continuously enrolled throughout the year based on plan enrollment period variables—namely, effective date and termination date. However, if claim adjusted effective/termination dates associated with that enrollment indicate 01/01/2020 and 01/31/2020, Inovalon includes medical claims for this individual *only for January 2020*, not for the entire year. As such, you should *not* misinterpret this to mean that the individual had zero medical claims from February through December 2020. Analogously, researchers should use Rx claim adjusted effective/termination dates to determine prescription claims data availability periods.

5.2 1M Representative Sample

The Harvard Inovalon Research Group at HMS created a representative sample of 1 million individuals. As this sample is not only representative but also follows the same structure as the entire dataset, researchers may find it useful to first examine the 1M sample to understand the data structure and assess the feasibility of their potential research ideas. The 1M sample is readily available on the HBS Grid as both text and Stata files. If

researchers wish to use these files, please reach out to Jinyeong Son.

5.3 Variable Length for Stata Users

Stata users should be aware that some numeric variables need to be imported as strings because Stata can only accommodate integers up to 16 digits, whereas some variables in Inovalon—specifically medical and prescription claims identifiers—contain up to 18 digits. Therefore, if these variables are loaded as numeric, Stata may incorrectly treat distinct claim identifiers as the same due to a loss of precision.

5.4 Output Export Process

We do not have a review process for exporting files (e.g., analysis results) outside the HBS Grid. Therefore, researchers may download their own outputs using SSH file transfer software, such as WinSCP. However, they must strictly adhere to all terms stipulated in the DUA, as well as other principles and policies governing research at Harvard.

6 Statistics

6.1 Covered Lives

We report the proportion of enrollee counts in Inovalon (*numerator*) to the corresponding enrollee counts reported in the survey (*denominator*) at the segment-by-state-by-year level. In these statistics, we consider only enrollees with plans that provide both medical and pharmacy benefits, while also accounting for the data availability periods discussed in Section 5.1. Using these enrollees, we construct the covered lives statistics as follows:

- **Baseline Numerator**: the number of enrollees (included in Inovalon) who have a health insurance plan as of *July 1st* in a given year,
- **Denominator**: the counterpart estimates from Kaiser Family Foundation (2024*a*), more specifically
  - Commercial = employer-sponsored plan enrollees + nongroup plan enrollees,
  - Medicaid = total Medicaid enrollees, and
  - Medicare = total Medicare enrollees.

Figure 1 shows the (nationwide) share of covered lives by segment from 2010 to 2020.

Using the results for 2019, Figure 2 demonstrates significant variation in the share of covered lives across states for all three segments. In particular, panel (b) shows that the proportions of covered lives for Medicaid *exceed* 100% in some states—which theoretically should not happen—suggesting a potential issue with the payer group information in Inovalon.

Next, we consider two alternative methods of constructing the numerators as follows:

- **Alternative Numerator 1**: the number of enrollees who have a health insurance plan on *any single day* in a given year,
- **Alternative Numerator 2**: the number of *full-year equivalent* enrollees who have a health insurance plan in a given year (e.g., two individuals with 6-month coverage would be counted as one).

In addition, one might consider that, since Inovalon does not include any Medicaid and Medicare FFS enrollees, it may be more reasonable to use the counts of Managed Medicaid enrollees and Medicare Advantage enrollees as the denominators.[7] We provide these supplemental statistics, along with other statistics using three different numerators, at the segment-by-state-by-year level in Excel format through the following links: Baseline, Alternative 1, and Alternative 2.

6.2 Continuous Enrollment Length

Next, we present statistics on continuous enrollment length—how many individuals we can follow over different time frames. Specifically, we first identify all enrollees observed as of 01/01/2015, and calculate their continuous enrollment days from that date onward. As before, we report the results separately by segment in Figure 3.

Focusing on commercially insured individuals in panel (a), we explain how to interpret the figures. As of 01/01/2015, 22,785,097 members had active commercial plans, with a mean continuous enrollment of 1,263 days. Among these members, 3,226,443 had a continuous enrollment of less than one year, with a mean continuous enrollment of 183 days. The analogous statistics—the numbers above the bars and the conditional mean days in the box—should be interpreted in the same manner. Analogous statistics at the state level can be found at the following link: Continuous.[8]

[7] The enrollee counts for Managed Medicaid and Medicare Advantage are from Kaiser Family Foundation (2024*c*) and Kaiser Family Foundation (2024*b*), respectively.

[8] Some state-level Medicare statistics on continuous days are masked due to small cell sizes.

6.3 Missing Data Rate by Variable

To give potential users a sense of which variables are relatively well-populated or not, Table 1 reports the missing rate for some key variables. It should be noted that we utilize the non-consolidated versions of the claims files in panels (c) and (d) below. Furthermore, we only consider clearly identifiable missing values (e.g., NULL or Unknown) in this calculation. However, some observations may contain invalid values, such as negative numbers for NDCs, which ideally should be treated as missing but are not considered so in this exercise.

7 Useful Resources

For those interested in learning more about Inovalon, we provide the following useful resources.

7.1 Resources

1. Data Dictionary: This data dictionary provides definitions for all the variables in each dataset within Inovalon.
2. User Guide: This Inovalon user guide provides a concise introduction to Inovalon, including topics such as the data generation process.
3. Inovalon Help Center: Inovalon operates a centralized platform (Help Center) to address users' questions and requests, while also providing additional resources and articles. You can sign up for an account using the link here and the registration code 3240965562.
4. Harvard Inovalon Data Issues and Solutions: At Harvard, Inovalon users, including both researchers and data scientists, collaborate to identify issues and propose potential solutions using this shared Google Spreadsheet.
5. Harvard ESTR: All administrative tasks related to the IRB process (i.e., DUA linkage, IRB, and Data Safety) can be completed here.
6. CITI Program: For the IRB and Data Safety application, the IRB office may require a specific CITI training certificate (e.g., Information Privacy Security). In such cases, you can log into this website through the organization and complete the required training.

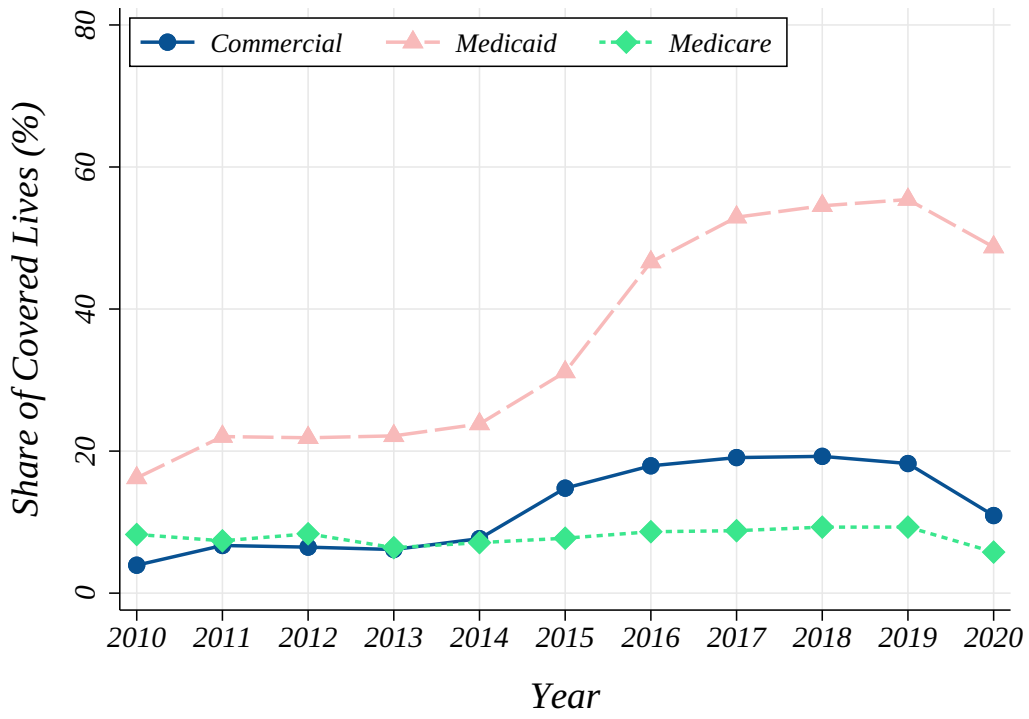
7.2 Contact

If you need assistance or additional information, please contact the following individuals:

- **Rachel Talentino**: rtalentino@hbs.edu (for IRB-related matters *only*)
- **Chirag Patel**: cpatel@hbs.edu
- **Jinyeong Son**: json@hbs.edu

8 Figures and Tables

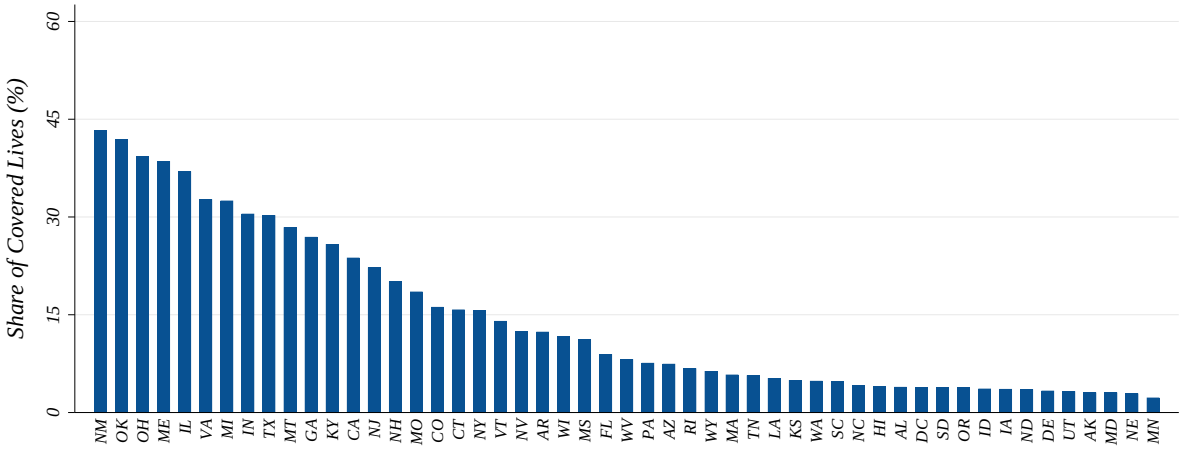
Figure 1: Share of Covered Lives by Insurance Type



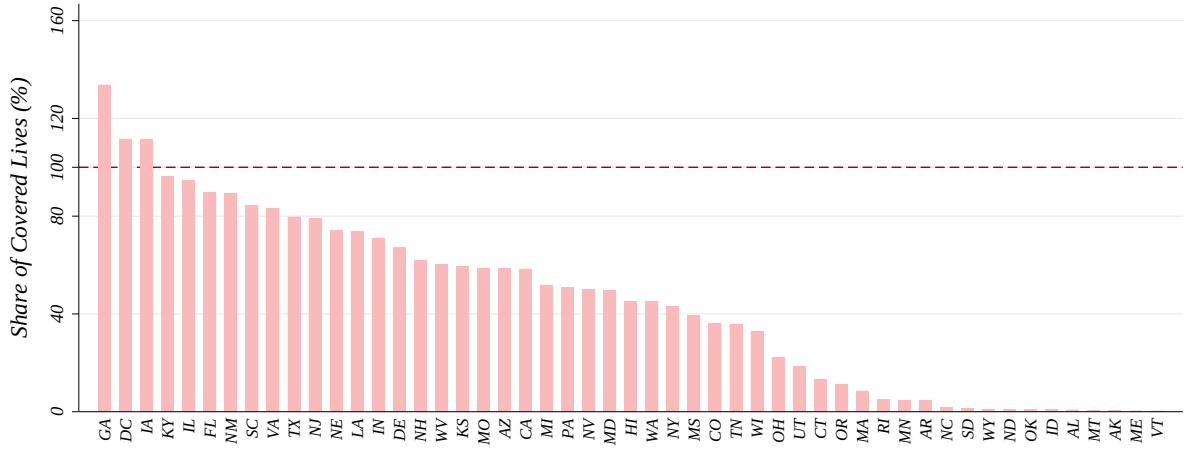
NOTES: One member may contribute more than once to the statistics if they are enrolled in more than one segment (i.e., dual eligibles) and satisfy the condition for inclusion in a numerator—e.g., enrolled in both commercial and Medicaid as of July 1 in a given year.

Figure 2: Variation in Share of Covered Lives Across States: 2019

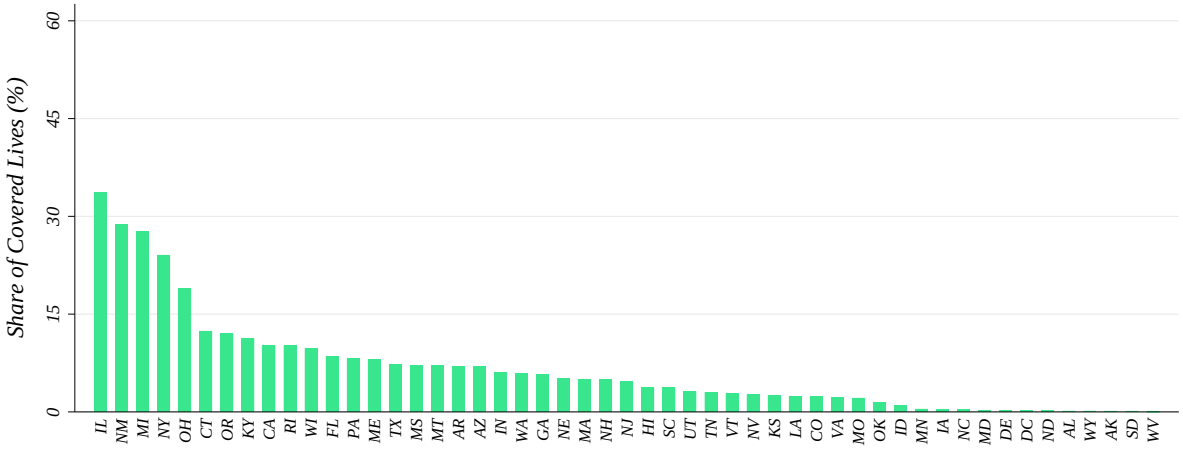
(a) Commercial



(b) Medicaid



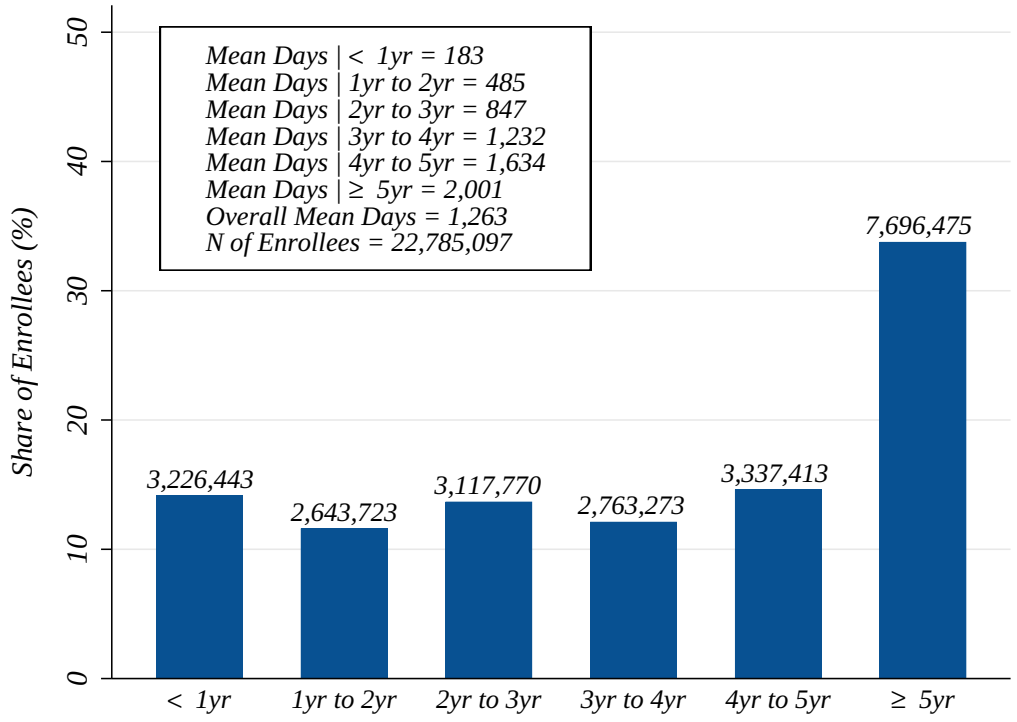
(c) Medicare



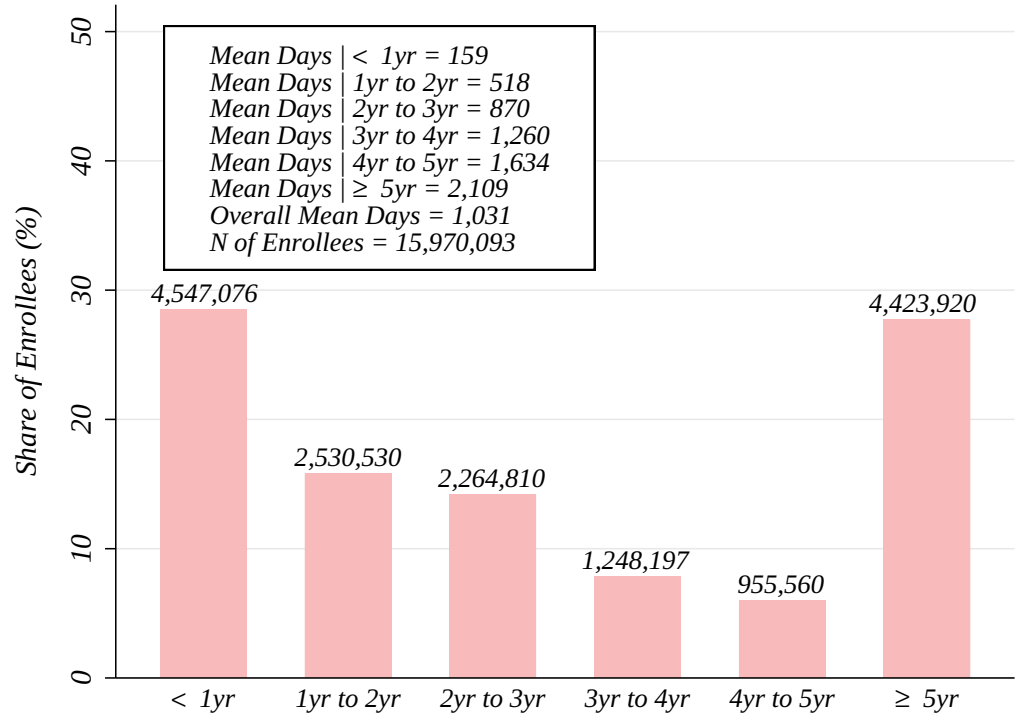
NOTES: One member may contribute more than once to the statistics if they are enrolled in more than one segment (i.e., dual eligibles) and satisfy the condition for inclusion in a numerator—e.g., enrolled in both commercial and Medicaid as of July 1 in 2019.

Figure 3: Continuous Enrollment Length Among Those Observed as 01/01/2015

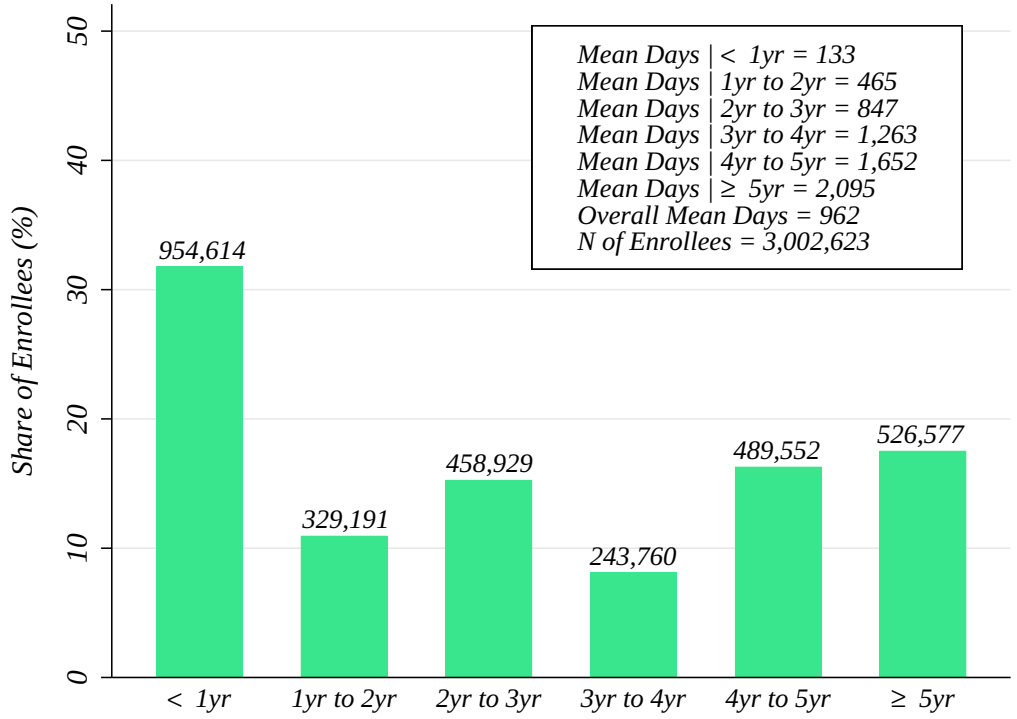
(a) Commercial



(b) Medicaid



(c) Medicare



NOTES: One member may contribute more than once to the statistics if they are enrolled in more than one segment (i.e., dual eligibles) and satisfy the condition for inclusion in a numerator—e.g., enrolled in both commercial and Medicaid as of January 1 in 2015.

Table 1: Share of Missing Values by Variable

Panel (a): Member

Variable:	Birth Year	Gender	State Code	ZIP Code	Race/ Ethnicity
Missing Rate (%)	1.60%	0.07%	1.03%	1.75%	55.57%

Panel (b): Enrollment

Variable:	Coverage Date	Payer Group	Payer Type	Product Type	Group Plan	ACA Indicator
Missing Rate (%)	0.00%	3.49%	3.49%	6.47%	82.56%	0.00%

Panel (c): Medical Claims

Variable:	Provider Info.	Claim Status	Service Date	Allowed Amount	Copay Amount	Paid Amount	Service Type
Missing Rate (%)	2.21%	0.11%	0.00%	64.66%	51.00%	61.05%	3.09%

Panel (d): Prescription Drug Claims

Variable:	Provider Info.	Claim Status	Fill Date/ NDC	Days Supplied	Allowed Amount	Copay Amount	Paid Amount
Missing Rate (%)	11.00%	0.01%	0.00%	0.00%	76.17%	50.59%	67.04%

Panel (e): Provider

Variable:	Provider Name	Org. Name	NPI	Practice Address	Billing Address	Taxonomy
Missing Rate (%)	23.98%	76.16%	13.18%	10.66%	11.96%	10.95%

NOTES: Missing organization names in the provider dataset can largely be filled by linking with the NPPES data.

References

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence From Patient Migration." *The Quarterly Journal of Economics*, 131(4): 1681–1726.

Kaiser Family Foundation. 2024*a*. "Health Insurance Coverage of the Total Population." https://www.kff.org/other/state-indicator/total-population (accessed August 2024).

Kaiser Family Foundation. 2024*b*. "Medicare Advantage: Total Enrollment, by Plan Type." https://www.kff.org/medicare/state-indicator/total-enrollment-by-plan-type (accessed August 2024).

Kaiser Family Foundation. 2024*c*. "Total Medicaid MCO Enrollment." https://www.kff.org/other/state-indicator/total-medicaid-mco-enrollment (accessed August 2024).