

Data Replication Package for “Competing under Information Heterogeneity: Evidence from Auto Insurance”

February 25, 2026

Overview

This replication package contains the code used for Cosconati et al., “Competing under Information Heterogeneity: Evidence from Auto Insurance.” We provide a collection of Stata and MATLAB scripts that construct the final estimation dataset and generate all tables and figures in the main text and Online Appendix. The data used in this project are confidential, and we provide general guidance on how interested researchers may seek access.

Data Availability Statement

Statement about Rights

The authors of the manuscript have proper access to and permission to use the data included in this manuscript.

Summary of Availability

The main dataset we use for this research is IPER (*Indagine sui Prezzi Effettivi RC Auto*), collected by IVASS, the Italian insurance supervising authority. These data are confidential and cannot be shared publicly, as they are protected under Italian privacy, confidentiality, and antitrust laws. Data may be obtained through agreements with IVASS. Certain additional secondary data files that are publicly available are nonetheless subject to dissemination restrictions on our part and are thus not provided in this replication package (as discussed further below).

Details on each Data Source

- The IPER dataset covers a nationally representative sample of matched insurer–insuree panel with rich information on observable risk factors, premiums, coverage, and contractual clauses. The data also include information on the frequency and severity of claims for each sampled consumer in each contract year. For our analysis, we restrict the sample to customers with tenure equal to zero in a given contract year for a given insurer within the province of Rome. These data are confidential and cannot be shared publicly.
- Some analyses in the Online Appendix use supplementary balance-sheet data from a module (n.17), submitted annually to IVASS by supervised insurance companies. These data include reported expenditures on customer service and claims liquidation for major insurers, which we use for external validation of our cost estimates. We are not authorized to share these data publicly.
- Additional analyses in the Online Appendix use supplementary datasets that include the number of employees at major insurers who list expertise in machine learning, data science, or artificial intelligence, the number of service centers each major insurer operates in the Rome metropolitan area, and a subset of pricing variables used by five major auto insurance companies. These data are publicly accessible from LinkedIn (which provides employee listings based on the insurer name and relevant search keywords) and the insurers’ websites (which list the locations of service centers and provide access to their online quoting systems), but we are not authorized to redistribute them directly.

Dataset List

The following list compiles all datasets needed to replicate the paper’s exhibits in the main text and Online Appendix. The datasets should be located in “01_data_prep/02_data”.

- “panel_roma.dta”: contains a sample of matched insurer–insuree panel data within the province of Rome, including information on observable risk factors, premiums, coverage, contractual clauses, and the frequency and severity of claims. This dataset is used to produce the main results in the paper. Details on how each table and figure is generated from this dataset are provided in the “Code and Instructions” section. Not provided.
- “variabili_pricing.xlsx”: contains the names of the pricing variables used by

five major auto insurance companies. This dataset is used to generate Table A.1 in the Online Appendix. Not provided.

- “external_measures.csv”: contains external validation measures for seven major insurers, used to generate Figure E.2 in the Online Appendix. Not provided.

Code and Instructions

Reproducing the results should take between 5 and 7 days on a standard Linux server with 10 CPU cores, 40 GB of RAM, and 8 GB of available disk space. The analyses were conducted using StataMP 15 with the installed package `ppmlhdfc` (available from SSC), and MATLAB R2023a. Below we describe all the code required to replicate the results in this paper, along with specific instructions on how to execute each component.

- Programs in “01_data_prep/01_code” generate the final dataset for estimation and the bootstrap samples, and should be executed in the following order:
 - “01_prep_claim_count.do”: This file prepares the data for recovering claim counts.
 - “02_recover_count.m”: This file recovers claim counts from a variable that indicates public accident records.
 - “03_gen_est_data”: This file generates final dataset for estimation.
 - “04_bootstrap_sample”: This file generates the bootstrap samples used to compute the bootstrap standard errors.
 - The raw data used for running these programs are stored in “01_data_prep/02_data”, and the outputs (including the final dataset for estimation, the bootstrap samples, and other intermediate datasets) are saved in “01_data_prep/03_working_data”.
- Programs in “02_estimation” generate the structural estimation results for three specifications. The programs in the following three folders are independent and can be executed in any order.
 - Programs in “02_estimation/01_main_estimation” generate the main structural estimation results. To produce the estimation results, conduct the model fit analysis, and compute bootstrap standard errors, execute the main file “main.m”. All other programs and functions in the same folder will be called automatically as needed. The estimation results are saved in “01_estimation_results” within the same directory.

- Programs in “02_estimation/02_out_sample” run the structural estimation using 80% of the data for estimation and the remaining 20% for validation. The main file to execute is “main.m”. All other programs and functions in the same folder will be called automatically as needed. The estimation results are saved in “01_estimation_results” within the same directory. The results produced in this folder are used to generate Table E.2 in the Online Appendix.
- Programs in “02_estimation/03_limit_consideration” run the structural estimation assuming consumers only consider top four insurers. The main file to execute is “main.m”. All other programs and functions in the same folder will be called automatically as needed. The estimation results are saved in “01_estimation_results” within the same directory. The results produced in this folder are used to generate results in Online Appendix G.
- Programs in “03_counterfactuals” generate the counterfactual results reported in the main text and the Online Appendix.
 - Programs in “01_baseline” recompute the equilibrium based on the model estimates. This serves as the baseline scenario. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
 - Programs in “02_efficiency_benchmark” solve the equilibrium in a setting where the true risk type of each consumer is observed by all firms. This serves as the efficiency benchmark. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
 - Programs in “03_centralized_bureau” solve the equilibrium in a setting where firms have equal access to aggregated risk information from a centralized bureau. This serves as our main policy experiment. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
 - Programs in “04_privacy” solve the equilibrium in a setting where the standard deviation of firms’ signal distributions is set to the largest level currently observed in the market. This approximates a privacy-regulation scenario in which firms are required to limit their use of consumer data. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.

- Programs in “05_information_alone” solve the equilibrium in a setting where all firms receive the aggregated risk evaluations from the centralized bureau while keeping their existing pricing strategies unchanged. This isolates the efficiency gains driven solely by more accurate information. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
- Programs in “06_off_equilibrium” solve the equilibrium in a setting where we improve Firm 1’s information precision to match the best in the market, while holding other firms’ pricing strategies fixed. This off-equilibrium scenario isolates the direct effect of better information on Firm 1’s pricing and performance. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
- Programs in “07_full_equilibrium” solve the equilibrium in a setting where we again improve Firm 1’s information precision, but now allow all other firms to adjust their pricing strategies in response. This setup captures both the direct effect and the general equilibrium effect through firm interactions in the market. The main file to execute is “main_cntf.m”. All other programs and functions in the same folder will be called automatically as needed.
- Programs in “08_summary” summarize the counterfactual results across all specifications described above. The main file to execute is “sum_cntf.m”. All other programs in the same folder will be called automatically as needed. The programs in the folders “01_baseline” through “07_full_equilibrium” are independent and can be executed in any order, except that the programs in “06_off_equilibrium” must be run after those in “01_baseline”. Programs in “08_summary” require the results from these seven folders and should be run only after those outputs have been generated.
- The folder “09_cntf_results” stores all counterfactual results across the specifications described above, as well as the summary results generated by the programs in “08_summary”.
- Programs in “04_tables” generate all tables reported in the main text and the Online Appendix. Table R.1 below summarizes, for each exhibit, the corresponding program and the location of its output. Table A.1 in the Online Appendix reports selected variables used by five major auto insurance companies. It can be generated directly from the dataset “variabili_pricing.xlsx”, so no additional code is required. The programs that generate the tables are independent and may be executed

in any sequence. For tables reporting estimation or counterfactual outcomes, the corresponding results must be obtained first before running the program that produces the table.

- Programs in “05_figures” generate all figures reported in the main text and the Online Appendix. Table R.2 below summarizes, for each exhibit, the corresponding program and the location of its output. The programs used to generate the figures are independent and may be executed in any order. “Figure_1.m” requires summary statistics produced by “Table_A3.do,” and therefore should be run only after the output of “Table_A3.do” has been generated and saved. In addition, for figures that report estimation or counterfactual outcomes, the corresponding results must be obtained before running the figure-generating program.

Table R.1: Summary of Programs and Outputs for Paper Tables

| Paper Exhibit | Program in “04_tables” Generating Output | Output File Produced in “04_tables /01_save_tables” |
|---|---|--|
| Main Text | | |
| Table 1: Estimates of demand-side parameters | Table_1.m | Table_1.txt |
| Table 2: Estimates of supply-side parameters | Table_2.m | Table_2.txt |
| Table 3: Counterfactual results: The impact of information policies | Table_3.m | Table_3.txt |
| Online Appendix | | |
| Table A.2: Regression of premiums on observable characteristics | Table_A2.do | Table_A2.xls |
| Table A.3: Summary statistics | Table_A3.do | Table_A3.xls |
| Table A.4: Regression of premiums on estimated risk type and observable characteristics | Table_A4.do | Table_A4.xls |
| Table A.5: Poisson regression of claim count on premium and observable characteristics | Table_A5.do | Table_A5.xls |
| Table A.6: Correlation coefficients between firms’ signal standard deviations, marginal costs, and claim processing efficiency | Table_A6.m | Table_A6.txt |
| Table A.7: Counterfactual results: Distributional effects on consumer surplus | Table_A7.m | Table_A7.txt |
| Table A.8: Percentage changes in counterfactual market outcomes: Comparison of information-alone and full-equilibrium outcomes under the centralized bureau policy. | Table_A8.m | Table_A8.txt |
| Table E.1: Model fit: Comparing data moments with simulated results using model estimates | Table_E1.m | Table_E1.txt |
| Table E.2: Out-of-sample model fit: Comparing data moments with simulated results using model estimates | Table_E2.m | Table_E2.txt |
| Table G.1: Comparing demand estimates under full or limited consideration set | Table_G1.m | Table_G1.txt |
| Table H.1: Counterfactual results: Off-equilibrium vs. equilibrium outcomes following an improvement in Firm 1’s information precision | Table_H1.m | Table_H1.txt |
| Table H.2: Percentage changes in profit: Off-equilibrium vs. equilibrium outcomes following an improvement in Firm 1’s information precision | Table_H2.m | Table_H2.txt |

Table R.2: Summary of Programs and Outputs for Paper Figures

| Paper Exhibit | Program in “05_figures” Generating Output | Output File Produced in “05_figures /01_save_figures” |
|--|--|--|
| Main Text | | |
| Figure 1: Heterogeneity across firms | Figure_1.m | Figure_1.eps |
| Figure 2: Regressing premiums on consumer risks | Figure_2.m | Figure_2.eps |
| Figure 3: Comparing offered and accepted price distributions for a firm using simulated data | Figure_3.m | Figure_3a.eps Figure_3b.eps |
| Figure 4: Who goes where: Rank correlation between information precision and average consumer risks across firms | Figure_4.m | Figure_4.eps |
| Figure 5: Differential impact on firms: Percentage change in profit relative to the baseline | Figure_5.m | Figure_5.eps |
| Figure 6: Sorting patterns: Average risk levels among consumers (measured in euros) within each firm under the baseline and three counterfactual scenarios | Figure_6.m | Figure_6.eps |
| Online Appendix | | |
| Figure E.1: Risk rating versus signal for Firm 1 | Figure_E1.m | Figure_E1.eps |
| Figure E.2: Rank correlations between model estimates and external measures for seven major insurers | Figure_E2.m | Figure_E2a.eps Figure_E2b.eps Figure_E2c.eps Figure_E2d.eps |
| Figure G.1: CDF of offered prices for the top four firms | Figure_G1.m | Figure_G1a.eps Figure_G1b.eps Figure_G1c.eps Figure_G1d.eps |

References

- [1] Istituto per la Vigilanza sulle Assicurazioni (IVASS). n.d. “IPER (*Indagine sui Prezzi Effettivi RC Auto*), 2013–2021 [dataset].” Last accessed at 2025-05-24.
- [2] Istituto per la Vigilanza sulle Assicurazioni (IVASS). n.d. “Supervisory Reporting Module No. 17: Supplementary Balance-Sheet Data.” Last accessed at 2025-06-09.
- [3] Cosconati, M., Y. Xin, F. Wu, and Y. Jin. 2025. “Supplementary Dataset on Insurer Workforce AI Expertise and Number of Service Centers.” Unpublished data. Last accessed at 2024-09-26.
- [4] Cosconati, M., Y. Xin, F. Wu, and Y. Jin. 2025. “Supplementary Dataset on a Subset of Pricing Variables of Five Major Auto Insurers.” Unpublished data. Last accessed at 2024-08-22.