CPB Netherlands Bureau for
Economic Policy Analysis

The Inequality of Labor and Health Risks: A Probability-Impact Perspective

This study decomposes the risk of labor and health shocks into their likelihood (shock probability) and impact (shock persistence). Using machine learning techniques and anonymous data on millions of Dutch people, we map out the entire likelihood and impact distributions. These turn out to be positively correlated, revealing an even greater risk inequality than was previously known.

We uncover multiple ways in which increased shock incidence and impact enhance one another. This stresses the vulnerability of groups where risks concentrate. Our findings facilitate proactive policies and further indicate their positive spillovers.

CPB Discussion Paper

Brinn Hekkelman, Pim Kastelein, Suzanne Vissers

March 2025

Doi: https://doi.org/10.34932/6j5x-q158

The Inequality of Labor and Health Risks: A Probability-Impact Perspective

Brinn Hekkelman[a], Pim Kastelein[a] & Suzanne Vissers[a,b]

March 26, 2025

Abstract

We decompose the risk of labor and health shocks along an extensive margin (probability of incidence) and intensive margin (impact given incidence, measured as shock persistence). The two margins turn out to be positively correlated, revealing a group of individuals that is particularly vulnerable and poorly resilient to adverse events. Leveraging machine learning techniques and Dutch administrative data, we show that shock persistence is predictable at the individual level and that it is unevenly distributed. By integrating our analysis with individual-level shock probability estimates from previous research, we reveal a triple-layered accumulation of risk: (1) ex-ante probabilities of shock incidence are correlated across the domains of labor and health; (2) higher ex-ante shock exposure is associated with greater shock persistence; and (3) experiencing multiple shocks amplifies the likelihood of persistent impacts. These insights can help inform the design of pro-active and unified policies that support vulnerable groups.

JEL classification: C53, H55, I10, J01, J64.

Keywords: risk inequality, labor shocks, health shocks, persistence, machine learning, prediction.

[a]CPB Netherlands Bureau for Economic Policy Analysis. Address: Bezuidenhoutseweg 30, 2594 AV, The Hague, The Netherlands.

[b]Corresponding author: s.j.p.l.vissers@cpb.nl.

We thank Johannes Spinnewijn and all participants of the CPB Research Seminar for their invaluable feedback. This project received funding from the Dutch Ministry of Social Affairs and Employment, for which we are grateful. We thank Jelmer de Jong for his exceptional research assistance.

1 Introduction

Life is inherently marked by challenges, and the extent to which adverse life events affect individuals varies. Job loss may be a temporary setback for some, while for others it can trigger a cascade of financial instability and long-term disadvantage (Athey et al., 2024). Similarly, health shocks—such as the onset of a chronic illness—can vary greatly in their consequences, depending on individual circumstances and resilience (Hoskins et al., 2024). Risk inequality in this context refers to the unequal distribution of both the exposure to and consequences of adverse life events, such as labor and health shocks, and forms the focus of this study.

Income and wealth inequality have traditionally been central concerns for economists and policymakers. We argue that risk inequality is a valuable addition to this discourse, as it directly impacts economic stability and can be at the root of income and wealth inequality. From a life-cycle perspective, repeated and prolonged exposure to shocks can hinder recovery, resulting in cumulative disadvantage over time, and thereby creating and perpetuating disparities. Groups disproportionately experiencing precarious employment are more likely to face negative *scarring* effects such as lower future income (Arulampalam, 2001; Nilsen and Reiso, 2011; Gregg and Tominey, 2005). In addition, health shocks, such as sudden hospitalization, have lasting spillover effects on employment and income (García-Gómez et al., 2013). Insight into how uncertainty is distributed prompts the consideration of solutions that mitigate risks for the most vulnerable rather than merely redistributing resources after inequalities emerge. However, since risks prove difficult to observe, there is a distinct lack of sufficient quantitative research explicating risk distributions and inequalities therein.

This paper advances the state of the art by quantifying risk distributions and uncovering risk inequalities across millions of people. We analyze both labor and health risks through their decomposition along two dimensions: an extensive margin and an intensive margin. Risk is conceptualized as the product of the probability of an adverse event (extensive margin) and the severity or persistence of its impact (intensive margin) (Kaplan and Garrick, 1981). By jointly analyzing probabilities and impacts, our research sheds light on the inequality of risk and uncovers the presence of particularly vulnerable groups, offering novel insights for policy making. When a higher probability of setbacks coincides with greater impact, the benefits of preventive measures become even more significant as they not only reduce the probability of immediate adversity, but also the persistence of its effect.

In this paper we estimate the intensive margin, which we consider to be the persistence of dependency on social benefits and lasting increased healthcare spending. We compile extensive administrative data on the Dutch population from 2013 to 2018, covering demographic and socioeconomic features, employment, income, wealth, and healthcare treatments. We train a machine learning model using a gradient-boosting algorithm and estimate the probability that a shock prevails at the individual level. Machine learning is tailored for this purpose, since it is optimized for predictive performance, capable of handling large datasets, and adept at discerning intricate interactions within the data, as described by Mullainathan and Spiess (2017). We then augment this intensive margin data with the extensive margin data from Cammeraat et al. (2023), which analyzed the predictability and concurrence of adverse events in the domains of labor and health.

The trained prediction models achieve high accuracy according to conventional metrics for assessing prediction performance. The predicted and realized recovery rates across the distribution are very close. To enhance the robustness of our findings, we conduct additional analyses. These involve analyzing the prediction performance for subsamples, extending the shock persistence window to two years, and refining the shock definitions to specific categories of social benefits and healthcare treatments.

Our results reveal significant heterogeneity in shock persistence across the sample. For the labor shock, two distinct groups can be observed: those with a high likelihood of prolonged dependence on social benefits, and those who are expected to return to work. This distinction holds when considering only unemployment benefits, while shocks in the form of dependence on social assistance or disability insurance show higher persistence. The distribution of health shock persistence exhibits a lesser degree of variation. This suggests that the process following a health shock tends to be less predictable. We find that mental health shocks are, on average, more persistent than physical health shocks, although the most persistent health shocks are of the physical type. In particular, the magnitude of the health shock measured in euros does not exhibit a significant correlation with persistence probabilities.

Next, the results reveal a positive correlation between the probability of shock exposure and the likelihood of the shock persisting over time. In other words, individuals with a higher ex-ante probability of facing setbacks are also expected to experience these setbacks for a longer duration. However, in the labor domain, the relationship between probability and impact does not increase monotonically. Specifically, individuals in the 5th to 20th percentile of persistence scores face a higher probability of experiencing a shock than those in the 20th to 50th percentile. This finding supports the hypothesis that individuals with flexible or precarious jobs are more likely to lose their employment but may also quickly find new work. For the rest of the distribution, a strong positive relationship is observed between the probability of shock exposure and its persistence, highlighting significant risk inequality for the majority of individuals.

In addition, we examine the impact of experiencing multiple shocks within the same year or in close succession. Our findings indicate that individuals who face an accumulation of setbacks are more likely to experience persistent shocks. Together, these results reveal a triple-layered accumulation of risk: (1) the ex-ante probability of experiencing a setback shows correlation across domains; (2) the ex-ante shock incidence probability is positively correlated with the probability of shock persistence; and (3) experiencing multiple shocks further increases the likelihood of shock persistence.

The multilevel accumulation of risks reveals a vulnerable group for whom intervention could yield positive outcomes. Examining the persistence distribution, we find that older, slightly wealthier individuals with fixed contracts are overrepresented in the high-persistency group. This challenges the dual labor market hypothesis, which suggests a division between a stable primary sector and an insecure secondary sector (Saint-Paul, 1996; Bentolila et al., 2019). With respect to the extensive margin, Cammeraat et al. (2023) found that individuals with a high probability of experiencing labor market shocks were often those on flexible contracts, suggesting segmented labor markets. However, our finding on the intensive margin that it is precisely individuals with fixed contracts who are overrepresented in the high-persistence group nuances this view of labor market segmentation.

In the case of health shocks, singles and single-parent families are overrepresented in the high persistence group. In addition, individuals in this group tend to be less wealthy and have lower education levels. These reduced dimensions of personal and human capital correlate with the persistence of negative outcomes, further exacerbating vulnerability.

Our results reveal a particularly vulnerable group with low resilience to adverse events. However, the fact that one can accurately predict the probability and impact of adverse events, and the significant heterogeneity in these predictions, creates a positive outlook for the identification and targeting of those most at risk. The variability in persistence scores indicates that support needs may differ after a shock occurs. Although some individuals may recover without additional assistance, others may require more intensive support, such as coaching and job search training. Furthermore, the positive correlation between probability and impact suggests that if policy successfully reduces the likelihood of a shock for the group with the highest probability of a shock, it will also significantly reduce the impact. This implication highlights the potential for broader benefits of proactive policy.

The paper proceeds as follows. Section 2 discusses relevant literature. Section 3 presents the data. Section 4 introduces the shock persistence definitions. Section 5 discusses the methodology and assessment of the predictability of the model. Section 6 discusses the distribution of shock persistence probabilities. Section 7 delves into the characteristics of those individuals with the highest and lowest persistence scores. Section 8 discusses the policy implications and section 9 concludes. Additional methodological details and supplementary analyses can be found in appendices A and B.

2 Related Literature

Income and wealth inequality are central topics in economic research, see, for example, (Atkinson and Bourguignon, 2000, 2014). Negative life events such as unemployment and illness play a crucial role in shaping these inequalities and are the topic of a growing literature. Unemployment can have long-term consequences, such as lower wages and an increased likelihood of repetitive unemployment (Arulampalam, 2001; Gregg and Tominey, 2005; Nilsen and Reiso, 2011). Furthermore, research by Kunaschk and Lang (2022), and Van Den Berg et al. (2023) finds that earnings losses following job displacement disproportionately affect already vulnerable workers, underscoring the need for targeted preventive policies.

In the health domain, Hoskins et al. (2024) examines individual-level differences in the persistence of health outcomes. The study finds substantial heterogeneity in the way health conditions persist over time, measured as the correlation between health in one period and health in future periods. This suggests that some individuals recover more slowly from health shocks than others. Danesh et al. (2024) explores the evolution of health inequality over the life cycle. The study reports that nearly half of the health disparity observed at age 70 is already evident by age 40, highlighting the value of preventive policies targeted at younger and middle-aged individuals.

Shocks in the labor and health domains are not isolated phenomena; rather, they tend to be interconnected, with negative events in one area exacerbating outcomes in the other. Cammeraat et al. (2023) investigates the predictability and concurrence of labor and health shocks, highlighting the extent to which adverse events cluster among certain groups. The literature has also established a strong link between health shocks and long-term spillovers. García-Gómez et al. (2013) analyze the long-term spillover effects of health shocks on unemployment. The other way around, Acevedo et al. (2020) studies the duration of unemployment on health outcomes. Acheampong and Opoku (2024) establish that a higher income inequality is associated with worse health outcomes.

The studies mentioned above focus on either the probability of experiencing a shock (extensive margin) or the persistence and severity of its impact (intensive margin). The innovation of our study lies in the integration of the two dimensions of risk, namely probability and impact (Kaplan and Garrick, 1981). By combining predictions of shock persistence with ex-ante probability estimates of shock occurrence, we offer a more comprehensive perspective on how risk is distributed across the population along its two dimensions.

In this regard, a recent study by Gregory et al. (2025) integrates both the occurrence and persistence of unemployment. The study identifies a group of workers that, when unemployed, quickly find a job and retain it, and a group of workers that experience prolonged unemployment spells and short job tenures. This classification is closely linked to the observed distribution of employment persistence and its relationship with ex-ante shock probabilities. Our paper shows that these types can be predicted in advance.

Our study also contributes to the social insurance literature, which concerns itself with the optimal design of insurancee schemes for adverse life events, such as health and unemployment shocks. When considering more generous insurance, such as extended or increased unemployment benefits, one should weigh the benefits against the costs, including moral hazard and adverse selection effects (Kolsrud and Spinnewijn, 2024; Chetty, 2009). Mueller and Spinnewijn (2023) shows that the probability of long-term unemployment is predictable and that the unemployment insurance coverage plays a limited role. This suggests that the benefits of unemployment insurance are larger than previously thought. Our finding of a positive correlation between the probability and impact of shocks further enhances the value of social insurance.

3 Data

This paper uses the same extensive data set used in Cammeraat et al. (2023), which is drawn from the administrative data infrastructure of Statistics Netherlands, the Dutch national statistical office. We combine data on all Dutch individuals across three domains:

1) demographic and socioeconomic characteristics, 2) employment, income, and wealth, and 3) healthcare treatments. This data is compiled from sources such as population registries, tax returns, and health insurance claims, and is accessible for scientific research through a remote access system, ensuring strict privacy compliance. The data modules can be linked using unique identifiers for individuals and households. Table A.1 provides further details on the type of information that is available to us in each of the three domains.

The dataset comprises approximately five hundred variables available annually, from 2013 to 2018. We construct a dynamic panel data set where the unit of observation is a person-year combination, see appendix A.4 for a detailed description. Demographic and socioeconomic characteristics are recorded at the beginning of each year, while employment information pertains to the highest earning job of that year. Healthcare treatments and expenses are aggregated by broad category throughout the year. Monetary values are adjusted to the 2015 price level. Missing values are not excluded from the analysis; instead, they are treated as informative by the machine learning models.

Once we have the raw panel data, we perform a sample selection. First, we exclude individuals younger than 25 or older than 60 years of age to focus on those most likely to be active in the labor force, mostly excluding students and pensioners. Then, we remove individuals with unknown or atypical household compositions, such as those in student housing or care homes. Importantly, we require complete data for all variables used in shock definitions for each person in a given year, ensuring accurate estimates for all shock definitions. To avoid a selection bias, this criterion does not apply to all other variables that are not used in the shock and persistence definitions. After sample selection the resulting data set contains approximately 12.6 million person-year observations.

Next, we select only those observations that have experienced a shock and we require that shock persistence, which will be defined in section 4 is computable. This results in two distinct samples: one for the labor shock and one for the health shock. These samples are no longer representative for the entire Dutch population, as the shock realizations concentrate

Table 1: Descriptive statistics of the total sample and the subsamples of individual-year observations exposed to a labor and health shock, respectively, and for which the persistence can be computed. The values are observed in the year before the shock year.

	Total sample		Labor shock		Health shock	
	Share	N	Share	N	Share	N
Observations	-	12,607,903	1.7%	212,477	2.9%	365,329
Gender						
Male	49%	6,226,127	45%	95,288	41%	151,482
Female	51%	6,381,776	55%	117,189	59%	213,847
Birth cohort						
1953 - 1962	24%	2,977,423	27%	57,237	33%	121,633
1963 - 1972	32%	4,005,691	29%	61,455	30%	108,958
1973 - 1982	25%	3,162,187	25%	53,153	21%	77,428
1983 - 1992	20%	2,462,602	19%	40,632	16%	57,310
Education level						
High	27%	3,415,686	25%	52,788	21%	74,967
Middle	26%	3,297,455	43%	92,373	28%	101,335
Low	13%	1,650,948	24%	51,923	19%	67,922
Unknown	34%	4,243,814	7%	15,393	33%	121,105
Household composition						
Couple with children	52%	6,549,619	43%	91,994	42%	154,992
Couple without children	23%	2,961,863	23%	50,331	28%	101,194
Single with children	7%	913,151	11%	22,994	9%	31,873
Single without children	17%	2,183,270	22%	47,158	21%	77,270
Migration background						
Dutch origins	77%	9,715,375	69%	145,961	77%	281,692
Child of migrant	8%	1,004,303	11%	22,509	8%	29,859
Migrant	15%	1,888,225	21%	44,007	15%	53,778
Home ownership						
Own house	68%	8,602,889	53%	113,152	61%	221,075
Rental house	21%	2,673,022	24%	50,339	22%	80,947
Rental house with rent allowance	11%	1,331,992	23%	48,986	17%	63,307
Wealth quintile						
First	27%	3,456,448	34%	72,249	28%	102,732
Second	15%	1,948,326	21%	45,477	20%	74,503
Third	19%	2,350,555	17%	35,799	18%	65,144
Fourth	21%	2,657,190	16%	35,004	19%	70,536
Fifth	17%	2,195,384	11%	23,948	14%	52,414

in certain demographic groups.

Table 1 presents a selection of characteristics for both the total sample and the subsamples of observations that have been exposed to the labor or health shock and meet the corresponding shock persistence computability requirement.[1] For both shocks, there is a slight over-representation of women and older people in the subsample. Furthermore, we see more people with a middle or lower education level, singles, people in a rental house and people from a lower wealth quintile[2] having been exposed to the labor and health shock. The distribution of migration background does not seem to vary for the total sample compared to the subsample of individuals who have been exposed to the health shock. For the labor shock, there is a higher incidence of shock exposure among persons with a migration background and their children.

4 Definitions of Shock Persistence

The shock in the labor domain is defined as the transition of an individual's main source of income to social benefits.[3] In the health domain, we consider the event of an increase in an individual's healthcare expenditures of at least 5,000 euros compared to the previous year.[4] These events correspond to the main shocks studied in Cammeraat et al. (2023). More background information on the variables underlying the shock and persistence definitions can be found in appendix A.2. In this section, we will pose the criteria for assessing the persistence of these shocks. These criteria should and do result in definitions one would

[1]For individuals who experienced a shock in 2018, the last year of the sample, we are unable to observe persistence due to lack of 2019 data. Furthermore, we lose individuals who no longer meet the sample selection criteria in the persistence year, such as those who turned 61 in the persistence year. Finally, we lose observations where the key variables required to determine shock persistence are missing, specifically, primary source of income and healthcare expenditures.

[2]These wealth quintiles are derived from the entire Dutch population. However, due to sample selection, the distribution of wealth in our study sample differs from that of the general population.

[3]This includes four types of benefits, namely unemployment benefits, disability/illness benefits, social assistance, and a set of other forms of social support. This last category includes benefits for young disabled people, older and partially disabled unemployed employees, and older and partially disabled former self-employed persons.

[4]This includes both physical healthcare and mental healthcare.

expect, i.e. a labor shock is persistent if reliance on social benefits persists and a health shock is persistent if a large part of healthcare expenditures persist.

4.1 Labor Shock Persistence

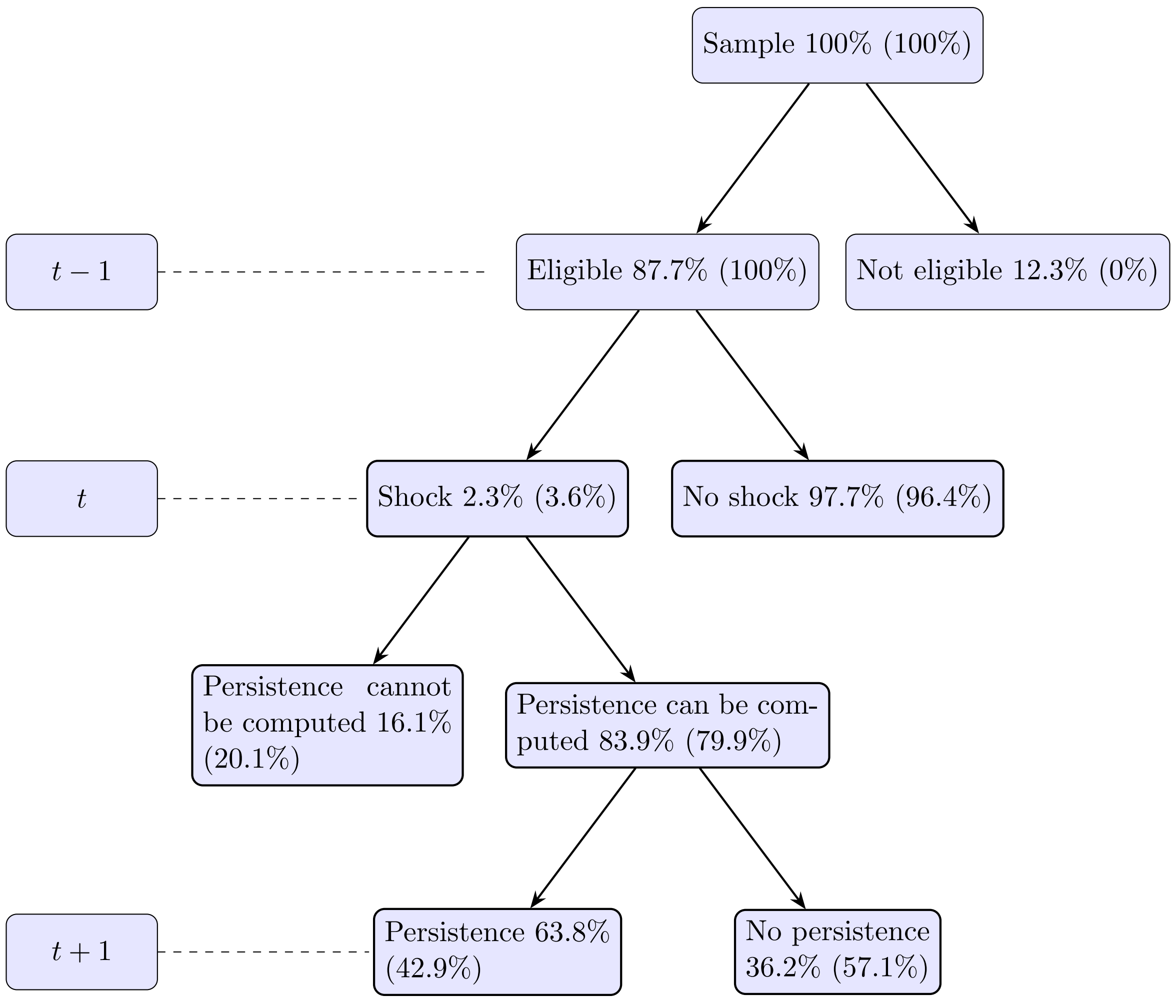
The general path of the shock and its potential persistence is given in figure 1. In each node, the first percentage relates to the labor shock, whereas the percentage in brackets refers to the health shock. Since the shock is defined as a status change, individuals who are already dependent on social benefits are excluded from the analysis. These individuals account for approximately 12.3% of the population. This split is visualized in the first step of the tree in figure 1. Each year, 2.3% of those eligible to experience the shock are affected. Of this 2.3%, about 16.1% of observations is discarded because the persistence cannot be computed.[5]

In the baseline analysis, we define the shock as persistent if, in the year following the initial event, social benefits remain the individual's primary source of income. Ultimately, of those observations that have been exposed to the labor shock and we are able to compute the persistence, the shock persists in 63.8% of the cases, see the last step in figure 1. If someone still relies on social benefits one year after the shock, then in 73.1% of the cases they rely on social benefits in the following year as well. Thus, while the time horizon of measuring persistence after one year is an arbitrary choice, it correlates strongly with persistence defined at longer time horizons.

We extend the analysis in two ways. First, we consider the labor shock as persistent only if the individual still relies on social benefits two years after the shock year. Second, we conduct a more detailed examination of the different types of social benefits. Rather than predicting whether the shock is persistent or not, we predict among the full range of states that the individual could transition into after the shock year. This includes retaining the same type of social benefits, switching to a different type, or not receiving any social benefits

[5] The main reason is that persistence cannot be computed for shocks occurring in the last year of our sample comprising 7 years of data.

Figure 1: Graphical representation of the timeline of the labor shock (health shock in brackets). Percentages are relative to preceding nodes. The composition of the sample in the first node can be found in the second and third column of table 1. In year $t-1$, the sample is divided into eligible and non-eligible individual-year observations. Among the eligible observations, a certain fraction experiences a shock in year t. Finally, we assess whether the shock is persistent in the following year $t+1$ for those individual-year observations where the shock persistence data is available.



anymore. Both supplementary analyses can be found in appendix B.

4.2 Health Shock Persistence

The health shock is defined as an increase of 5,000 euros or more in an individual's healthcare expenditures within a single year. Note that there is no eligibility condition for this shock, as is depicted in the first step of figure 1. The health shock occurs in about 3.6% of the person-year observations in the data set. We define the shock as persistent if the increase in healthcare expenditures from the shock-year has decreased by no more than 80% (i.e. expenditure increase has not decreased to less than 20%) in the following year.[6] For example, if a person's healthcare expenditures increased by 10,000 euros in year t, the shock is considered persistent if their expenditures in year $t+1$ have not decreased by more than 8,000 euros. We lose approximately 20.1% of the sample when we require the persistence condition to be computable (again a large fraction being accounted for by the last year in our data). Of the remaining observations, we observe shock persistence in 42.9% of the cases. If one's increase in healthcare expenditures do not revert by at least 80% after one year, then in 56.2% of the cases they also do not revert after two years. The health shock persistence measured after one year thus correlates strongly with persistence measured at longer time horizons.

Analogous to the extensions of the labor shock, we extended the analysis in the health domain in two ways. First, we consider the shock to be persistent only if the healthcare expenditure decrease compared to the shock year does not exceed 80% after two years. Second, in appendix B.3 we perform a multi-class analysis where we zoom in on the type of healthcare: physical or mental.

5 Predictability of Shock Persistence

In this section we present the machine learning methodology we apply to estimate persistence probabilities. Machine learning techniques are ideally tailored for this purpose as they are able to assign probabilities to situations based on past realizations for similar situations. Extracting underlying probabilities from observed realizations is one of the main challenges in risk inequality research. Evaluation of the models' performance shows that the uncov-

[6]The choice of 80% decrease, at which approximately half of the shocks persist, provides a practical level that enhances the model's performance and reliability. The fraction of persistent shocks scales approximately linearly from 30% to 75% when varying the threshold from 50% to 100%.

ered probabilities align with aggregated realizations to a high degree. This means that our probability estimates constitute an accurate representation of real persistence probabilities.

5.1 Methodology

For each individual i that had a shock in year t we observe whether that shock persisted in year $t+1$. This realization of persistence in year $t+1$ is a binary variable: its value is 1 if the shock persisted and its value is 0 if it did not. What this binary realization fails to accurately capture is the underlying probability that individuals' shocks will persist in time. To assess the ex-ante impact of a shock in terms of its persistence, it is exactly this underlying probability that we are interested in.

To uncover the underlying probabilities of interest, we apply machine learning methods similar to Cammeraat et al. (2023) using the *R* package *LightGBM*. Our goal is to predict, for each individual i with a shock in year t, the probability $p_{i,t+1}$ that the shock will persist into year $t+1$. Using data from the shock year and prior years, we train a machine learning model to estimate shock persistence, assigning scores between 0 and 1. If accurate and calibrated, these scores represent estimates of the true persistence probabilities. For both shocks, we train a gradient boosted tree model on a train set, and then subsequently have them make persistence predictions on a test set that we use in our analyses. This prevents data leakage and ensures fair performance. For the full set of model parameters, see appendix A.3. For the full description of the data pipeline, see appendix A.4. For some insight into variable importance, see appendix A.5.

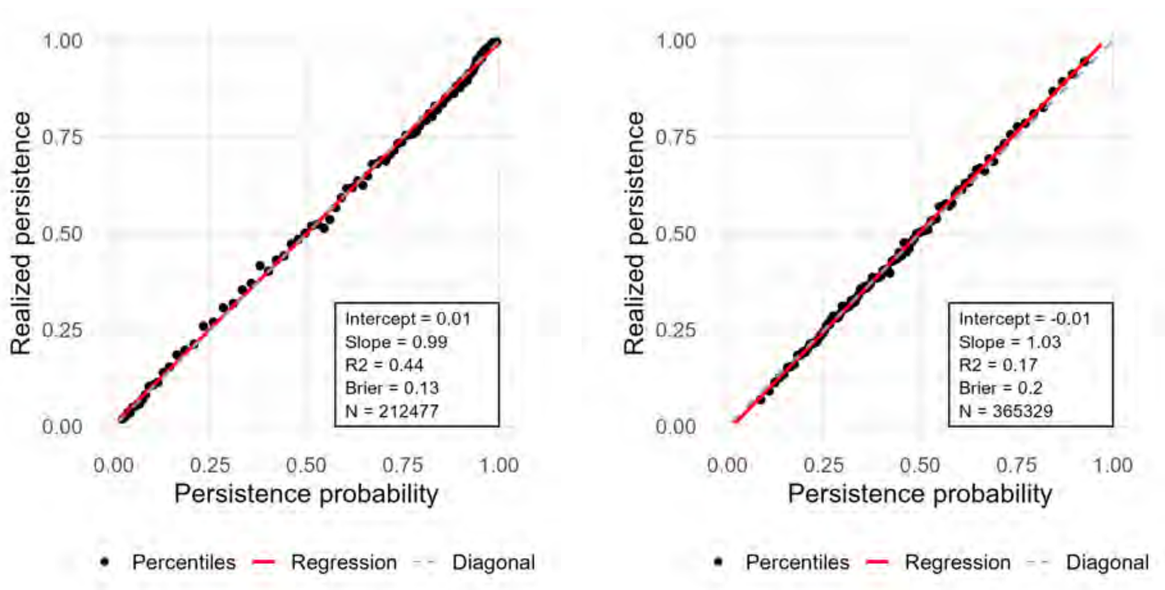
Analogous to predicting shock persistence in the year $t+1$ we also analyzed shock persistence in the year $t+2$. This is a form of increased impact, as the shock persists for longer. Nevertheless, we find similar results in terms of predictability, prevalence, and distributions. These can be found in appendix B.2.

5.2 Performance Evaluation

Since true probabilities are unobservable, we assess prediction performance by comparing our probability estimates to realization prevalence at an aggregated level. Figure 2 shows a regression of the persistence realizations on probability estimates for both shocks, along with percentile bins of relative probability estimates and their realization prevalence. If the model is unbiased, the regression should align with the 45° line, which is observed with intercepts near 0 and slopes near 1.

Figure 2: Regression of persistence realizations on persistence probability estimates for the labor shock (panel a) and health shock (panel b).

(a) Labor shock (b) Health shock



The regression line aggregates over the entire dataset, but we further evaluate predictions by dividing the test set into percentile bins of probability estimates. For all observations per bin, we compare the mean probability estimate to the realization prevalence. An unbiased model should produce estimates close to the prevalence, which is confirmed by the scatter points in figure 2. Also noted in figure 2 are the R-squared and Brier scores.[7] The higher

[7]The R-squared indicates how much of the variation in the actual realizations is predicted by the ex-ante probability estimates, higher meaning more predictable. For reference, Mueller and Spinnewijn (2023)

R-squared indicates that labor shock persistence is more predictable than health shock persistence, while the lower Brier score shows that the labor model is better calibrated, with its probability estimates more accurately reflecting the actual outcomes.

We find that the performance of our prediction model is similar when assessed for various population sub-samples. For example, values similar to those in figure 2 emerge when the sample is split up by person characteristics such as gender, age, education level, household composition, migration background, home ownership and wealth level. It thus appears that our prediction model performs similarly for different groups of individuals. The prediction performance is also comparable regardless of whether someone had a relatively low or high ex-ante probability of receiving the shock and regardless of the specific type of social benefits or healthcare treatment someone receives. These results can be found in appendix A.6 along with additional performance evaluation metrics.

6 Distribution of Shock Persistence

In this section, we first quantify the distribution of shock persistence across the population, highlighting the variability in how individuals experience the lasting impacts of shocks. For labor we find a clear bimodal distribution while health is more concentrated. We then augment these findings with ex-ante shock probabilities from Cammeraat et al. (2023) to reveal a positive correlation between the likelihood of shock incidence and its subsequent persistence. Finally, we analyze the correlation between shock persistence and the occurrence of other shocks, yielding a quantification of how multiple adverse events amplify vulnerability.

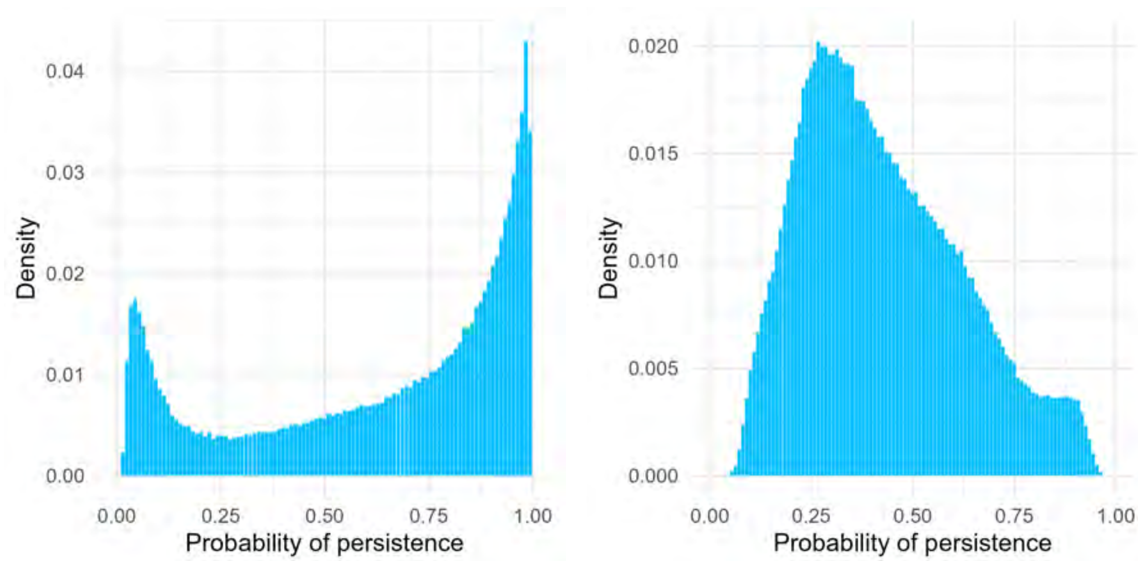
obtain an R-squared of 0.136 when predicting job finding rates with similar methodology. The Brier score measures the average difference between probability estimates an realizations, with lower scores indicating better calibration.

6.1 Probability Distributions

Figure 3 shows the distribution of the predicted probabilities for the persistence of both shocks considered. The bimodal distribution in figure 3(a) highlights a clear divergence in the persistence of the labor shock. That is, it indicates the presence of two distinct groups: one for whom the shock is likely to persist and another for whom the likelihood of no longer relying on social benefits is high.

Figure 3: The histograms display the density of persistence probabilities for the labor shock (panel a) and the health shock (panel b).

(a) Labor shock (b) Health shock



One possible explanation for the bimodal distribution is that it arises from a combination of two or more unimodal distributions corresponding to different types of social benefits. Appendix B.1.2 discusses the results for the different types of underlying shock types. Individuals reliant on disability insurance or social assistance typically exhibit high persistence scores, whereas the distribution for unemployment benefit shocks remains bimodal.

Figure 3(b) reveals an opposite pattern for the health shock. The unimodal distribution and thin tails suggest that only in a very few cases the prediction model is almost certain of whether the shock will persist or not. This points to a greater degree of uncertainty and randomness surrounding the persistence of health shocks, and aligns with the lower predictive power reported in section 5.2.

We may expect some heterogeneity based on the magnitude of the initial health shock. However, as shown in appendix B.1.1, the magnitude of the health expenditures shock does not exhibit a strong relation with the persistence probability. In appendix B.1.2, we distinguish the health shock into a mental health shock or a physical health shock, based on the highest cost for either category. The analysis shows that the physical health shock is typically associated with a lower probability of persistence than the mental health shock. However, the highest probabilities of persistence are observed for physical health shocks.

The reduced predictive power of persistence of the health shock compared to the labor shock is in line with the lower prediction power for the occurrence of the health shock compared to the labor shock in Cammeraat et al. (2023). While labor shocks and their persistence exhibit a clearer separation in risk profiles, the prediction of health shocks is hindered by the higher level of uncertainty, leading to a more diffuse probability distribution.

6.2 Shock Probability versus Shock Persistence

We continue by exploring the link between the ex-ante probability of being exposed to a shock and the probability of the shock being persistent. Figure 4 illustrates the relation between the probability of the shock being persistent in percentiles (x-axis) and the average ex-ante probability of receiving the shock (y-axis). For both shocks, we observe a generally positive relation. That is, the individuals who are most likely to experience a persistent shock, are also those who were ex-ante most likely to be exposed to the shock. appendix B.1.3 discusses the relation between the shock probability and shock persistence for the different underlying shock types.

6.2.1 Labor shock

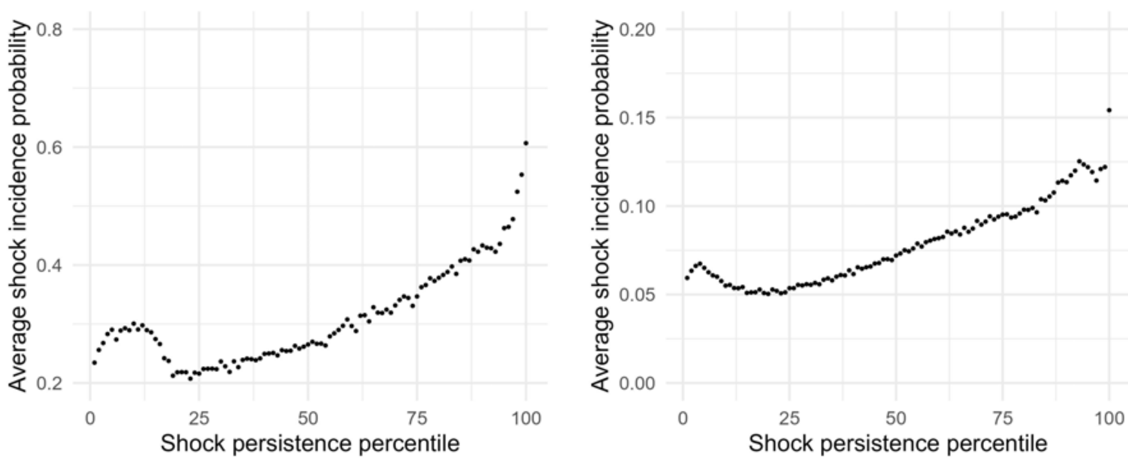
The link between the probability of unemployment and its persistence can be hypothesized in two directions. On the one hand, individuals with a high likelihood of unemployment may struggle to find new employment quickly due to skill mismatches, a precarious work history, or weaker professional networks. However, those in flexible or temporary roles may be more prepared for job transitions and proactive in seeking new opportunities. Both outcomes depend on individual circumstances and labor market conditions.

The average ex-ante labor shock incidence probability is mostly increasing in the shock persistence percentiles, except for a non-monotonic bump in the left tail, as can be seen in figure 4(a). This bump could be explained by the fact that this group partly consists of young people with a flexible work contract. This group has a high shock incidence probability but is also likely to find work again. The range of the average ex-ante shock probability spans from approximately 0.2 to 0.6.[8] This positive relation between probability and impact marks the existence of a vulnerable group that has both a high probability of relying on social benefits and a high probability of being dependent for a continued period of time.

[8]The distribution of the probabilities found by Cammeraat et al. (2023) ranged from close to 0 to ~ 0.7. A large group was found to have a near-zero risk of facing the labor shock, and a much smaller group had a significant risk. Logically, in the sample of this study of shock persistence, there is an over-representation of people with a higher shock incidence probability.

Figure 4: The figures show the relationship between ex-post persistence probabilities and ex-ante shock incidence probabilities. The x-axis represents percentiles of estimated persistence probabilities, while the y-axis shows the average shock incidence probability for individuals in each percentile.

(a) Labor shock (b) Health shock



6.2.2 Health shock

We expect a positive relationship between the probability of experiencing a health shock and its subsequent impact, as individuals with higher risk factors - such as preexisting health conditions - are generally more vulnerable to severe consequences. However, this relationship may not be entirely monotonic. For instance, younger individuals might be at higher risk due to more risky behavior but may experience less severe impacts if they are otherwise healthy.

The ex-ante health shock incidence probability is also roughly increasing in the shock persistence percentiles. Note that the range of the shock incidence probabilities is smaller than in the labor domain, ranging from approximately 0.05 to 0.15. The underlying shock incidence probability percentiles range from 0.01 to 0.25. Hence, the more extreme values are averaged out, but we still observe a strong heterogeneity. Nevertheless, significant heterogeneity remains evident.

6.3 Conditional Probabilities

Beyond the correlation between shock incidence and persistence probabilities, we are similarly interested in the correlation between past shock realizations in other domains and shock persistence. Cammeraat et al. (2023) showed that conditional on having experienced one shock, the probability of being exposed to a shock in another domain in the next year increases significantly. Apparently there is a compounding effect to suffering multiple coinciding shocks.

We therefore investigate how suffering coinciding shocks affects the probability of persistence of the main shock. To do this, we look at individuals that realized one of our two main shocks in year t and also realized a coinciding shock from another domain in the same year t or the previous year $t-1$. We then look at how the shock persistence of the main shock differs for these individuals relative to all individuals who realized that shock. The restriction of realizing both shocks in more or less the same year means only a rare few cases in our data are available for analysis. However, since we have such a large number of individuals available in our initial data set, we still have between 1,000 and 20,000 observations per coinciding shock to base our analyses on.

Table 2 lists the persistence outcomes for our main shocks conditional on coinciding shocks.[9] We immediately observe an increase of labor shock persistence to that of the highest percentiles in the unconditional case as shown by figure 3. This indicates that suffering a coinciding health shock greatly increases the probability of the labor shock persisting. Moreover, the observed correlation holds for differing types of health shocks. Intensive care admittance often indicates a serious condition, and only few individuals suffer this shock. Starting mental health medication, on the other hand, is much more common. Still, both shocks increase the probability of subsequent labor shock persistence to around 80%.

[9]The additional shocks mentioned in table 2 are defined in line with Cammeraat et al. (2023) as follows. *Intensive care:* individual spends at least one day on intensive care and none in the year before, 0.3% prevalence in total population. *Mental health medication:* individual starts taking antidepressants, antipsychotics or sedatives, 2.3% prevalence in total population. *Problematic debt:* individual starts defaulting on health insurance premium payments, 0.5% prevalence in total population.

Table 2: Shock persistence conditional on facing another coinciding shock in the same year t or the previous year $t-1$. N denotes the number of individuals in our data that suffered the specific combination of shocks. The prevalence expresses for which portion of these individuals the main shock persisted in year $t+1$, and the probability is the model's corresponding average estimate. The bold lines illustrate baseline results without conditioning on the presence of a coinciding shock.

Main shock	Coinciding shock	N	Prevalence	Probability
	-	**212,477**	**0.637**	**0.639**
	Health (t)	8,411	0.730	0.733
	Health ($t-1$)	8,750	0.694	0.701
	Intensive care (t)	1,177	0.836	0.844
Labor	Intensive care ($t-1$)	1,127	0.833	0.834
	Mental health meds (t)	11,333	0.794	0.782
	Mental health meds ($t-1$)	11,888	0.758	0.759
	Problematic debt (t)	3,871	0.664	0.659
	Problematic debt ($t-1$)	3,751	0.692	0.683
	-	**365,329**	**0.429**	**0.427**
	Labor (t)	17,546	0.448	0.445
Health	Labor ($t-1$)	13,020	0.462	0.460
	Problematic debt (t)	2,861	0.439	0.440
	Problematic debt ($t-1$)	3,308	0.455	0.451

The correlation between health shock persistence and coinciding shocks in the labor domain is less pronounced. This is line with the lower predictability of health shocks in general. Apparently, health shocks and their subsequent persistence are less correlated to observable variables in the years prior. It is not surprising that we again find this distinction between the two domains: developments in the health domain are harder to see coming based on the type of administrative data that we have access to.

It is striking that the models' probability estimates accurately reflect the realized prevalences, considering that the models were not trained to predict conditional shock persistence outcomes. Apparently, these correlations were picked up during training to such an extent as to be clearly observable when singled out.

7 Personal Characteristics

In this section, we analyze the characteristics of individuals most likely to experience a persistent shock. This paints a picture of mostly confirmed expectations: women, foreigners, elderly, lowly educated, etc. are overrepresented in the top quintile. However, these characteristics are found across the distribution and therefore cannot by themselves serve as sufficient information for targeting. Interesting to note is that persons with fixed contracts are much more resilient to labor shock persistence than those without.

Our focus is on the features of individuals in the year prior to the shock.[10] We divide the sample into three groups based on the probability of shock persistence: the bottom 20% (those with the lowest probability of shock persistence), the middle 60%, and the top 20% (those with the highest probability of shock persistence). For a selection of personal characteristics, we examine their distribution across these three groups. The results are presented in figure 5 and figure 6 for the labor and health shock, respectively.

Gender. As reported in table 1, more women than men experienced the labor shock, with 55% of the affected individuals being women and 45% men. Additionally, in the top 20% of individuals wil the highest probability of experiencing a persistent labor shock, approximately 62% are women and 38% are men. In the health domain, the health shocks are also more common among women than men, with a distribution of 59% to 41%. When we then look at the persistence, we see that both the lower and upper tail of the distribution contain a relatively high proportion of women.

Country of origin. Figure 5(b) indicates a slight overrepresentation of individuals born outside the Netherlands in the middle group of labor shock persistence. In contrast, individuals of Dutch origin are relatively more concentrated in both the low and high probability groups for labor shock persistence. Regarding the health shock, there is a slight overrepresentation of individuals born outside the Netherlands within the top 20% of those with the highest probability of shock persistence.

[10] For many features, these values will be equal or similar to their values in the subsequent years.

Birth cohort. Figure 5(c) shows that the labor shock is most persistent for the older generations, whereas the younger individuals are more often expected to find a job again. Surprisingly, figure 6(c) shows a non-monotonic relationship between the risk groups and the share of the oldest generation.

Education level. The likelihood of relying on social benefits for an extended period appears to be negatively correlated with education level. As shown in figure 5(d), individuals with higher education are relatively underrepresented in the top 20% persistence group for labor shocks. A similar pattern is observed for health shocks, with higher education levels associated with lower probabilities of shock persistence.[11]

Employment relationship. For the labor shock, we examine the nature of employment relationships in the year preceding the shock year t. Notably, individuals with fixed employment contracts are overrepresented in the high-persistence group. While Cammeraat et al. (2023) demonstrated that those most likely to experience labor shocks are typically on flexible contracts, our findings suggest that, once unemployed, individuals with fixed contracts experience more persistent impacts. This may be associated with factors such as age and the accumulation of more extensive unemployment benefit entitlements.

Wealth. Figure 5(f) illustrates the relationship between an individual's wealth quintile and their position in the persistence group. Notably, there is a positive association between wealth and the likelihood of experiencing a persistent labor shock. This relationship may be influenced by age, as older individuals, who tend to have accumulated more wealth, are also more likely to remain on social benefits. In the case of health shocks, as shown in figure 6(f), the second wealth quintile is overrepresented in the top 20% persistence group, while the wealthiest quintile is underrepresented.

[11]The relatively high number of missing observations is primarily due to gaps in administrative data on education levels, which are most prevalent among older individuals and immigrants.

Figure 5: Group analysis for persistence of labor shock by a selection of personal and household characteristics.

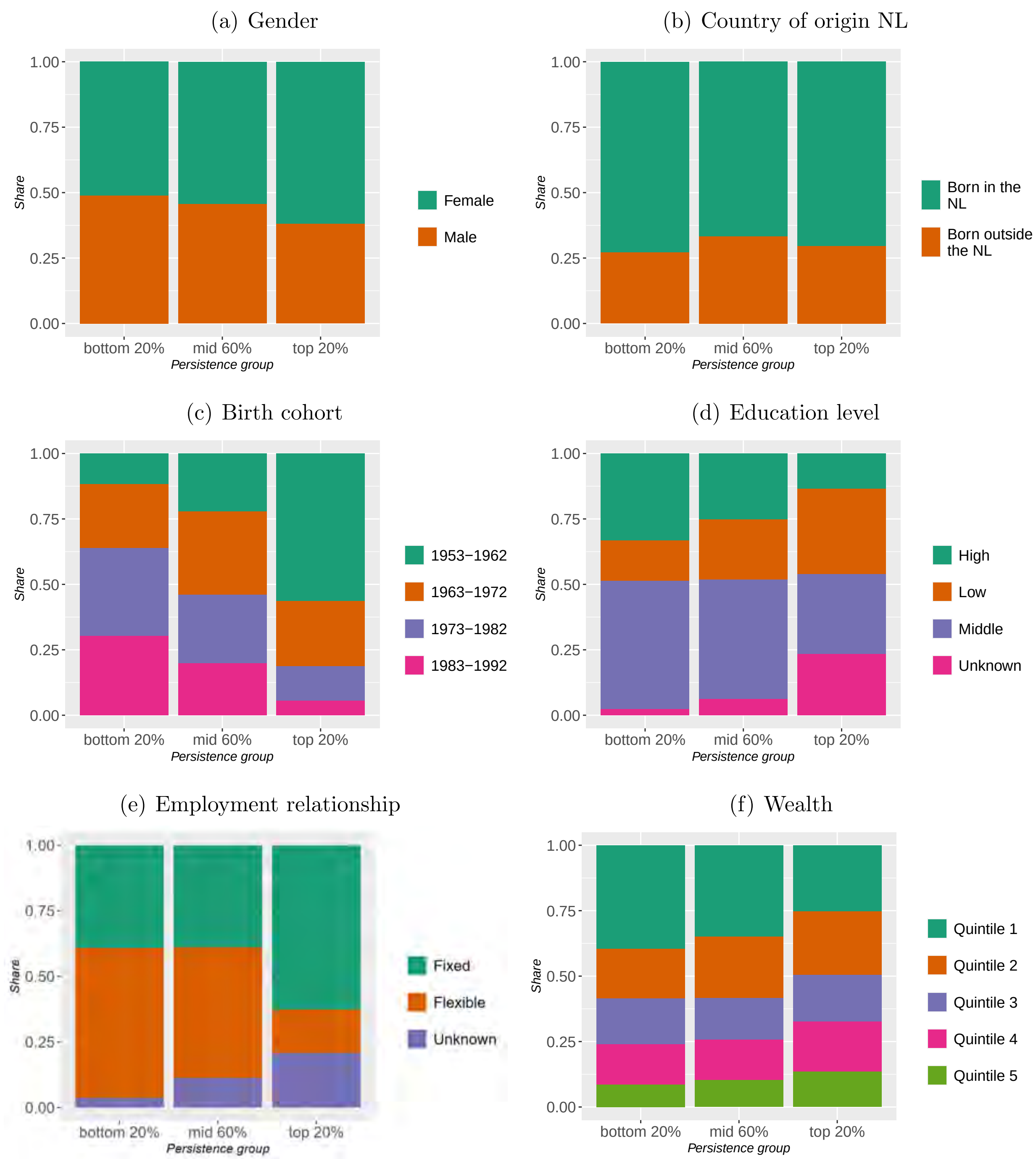
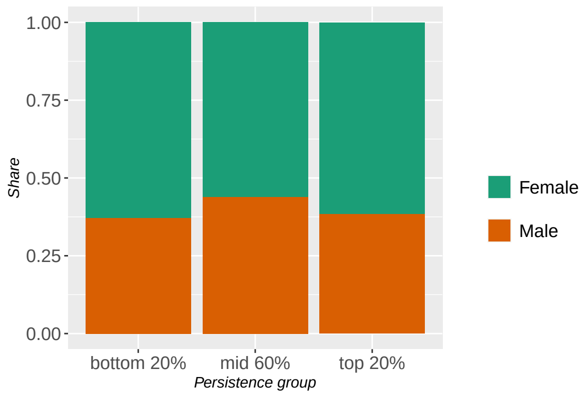
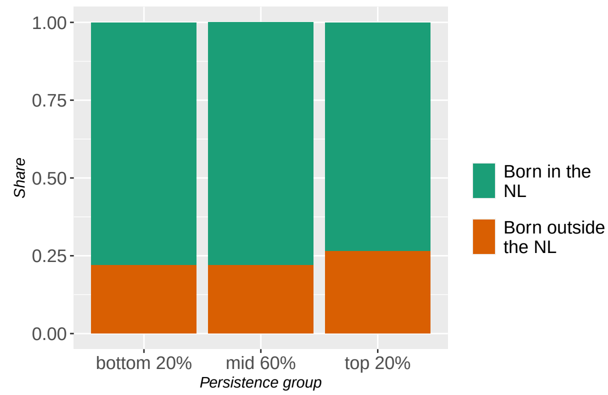


Figure 6: Group analysis for persistence of health shock by a selection of personal and household characteristics.

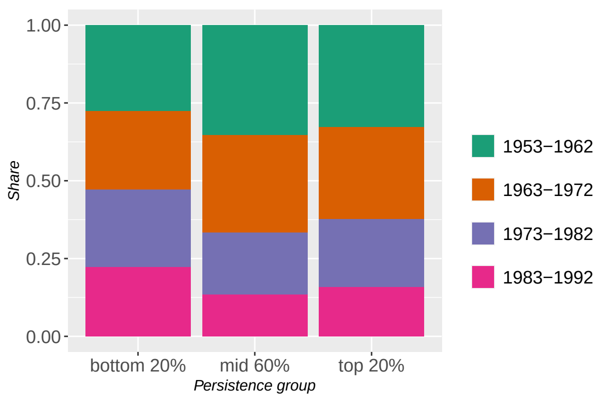
(a) Gender



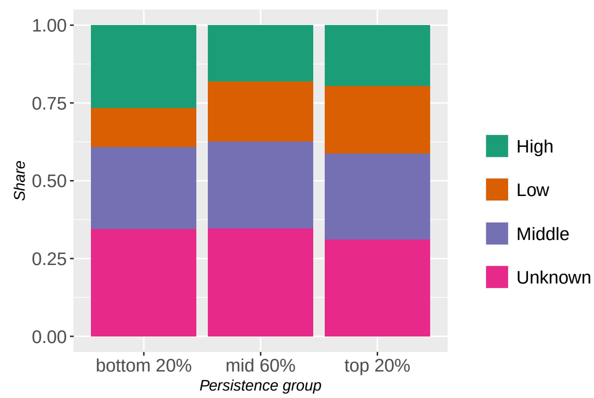
(b) Country of origin NL



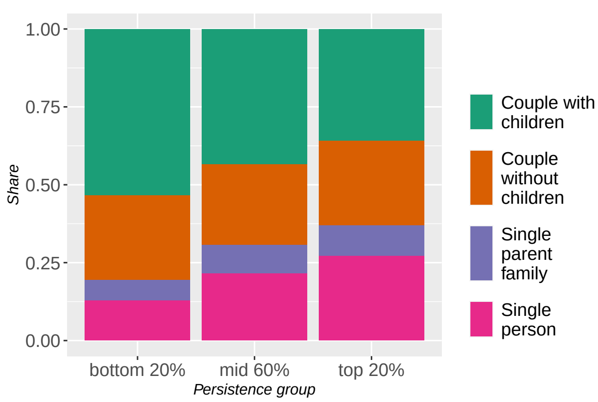
(c) Birth cohort



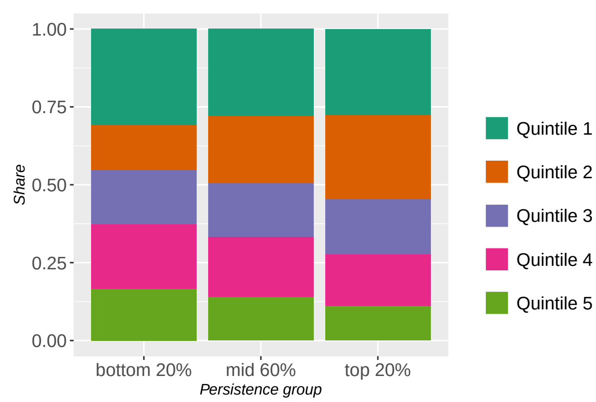
(d) Education level



(e) Household composition



(f) Wealth



8 Policy Implications

Traditionally, income and wealth inequality have been central concerns for policymakers. We argue that risk inequality—the unequal distribution of exposure to and consequences of adverse events— is a valuable addition to this discourse, as it directly impacts economic stability and can be at the root of income and wealth inequality. This paper demonstrates that risk inequality is not only unevenly distributed in terms of the probability of experiencing labor and health shocks, but also in their persistence, highlighting a critical dimension of socioeconomic inequality.

Not only are the two dimensions of risk unequally distributed, but risk also tends to accumulate in distinct ways. This study highlights a triple-layered accumulation of risk. First, Cammeraat et al. (2023) demonstrated that the likelihood of encountering setbacks is correlated across different domains. Second, our findings reveal a positive correlation between shock occurrence and shock persistence, indicating that higher initial probabilities are associated with longer-lasting impacts. Third, we show that individuals who have experienced multiple shocks simultaneously or in close succession have a higher probability of shock persistence, underscoring the compounding effect of concurrent adversities.

Our analysis identifies a particularly vulnerable group exposed to triple-layered risk accumulation. Encouragingly, this study also demonstrates the potential for accurately identifying individuals within this high-risk group. Not only can we reliably predict who is likely to experience a persistent shock, but we also observe substantial heterogeneity in the probabilities of shock persistence, which further enables targeted identification and intervention for those most at risk. The positive correlation between probability and impact further implies that if a policy measure can successfully target those at high risk of occurring a shock, comes with the broader benefit of preventing those events with the highest impact of the shock.

While our analysis captures structural patterns in labor market and health shocks, it also reflects the influence of institutional incentives. The observed persistence of social benefit dependence is not solely a consequence of individual characteristics or labor market conditions but is, to some extent, shaped by the design of the current system. The structure of benefits and eligibility criteria can create incentives that discourage rapid re-entry into the labor market. In particular, individuals who have accumulated substantial unemployment benefit rights may face weaker incentives to return to work quickly. This suggests that policy design plays a crucial role in reinforcing or mitigating persistent labor market disengagement.

The next step for policymakers is to translate the predictions into concrete actions, but several challenges remain. Key uncertainties include the effectiveness of different policy measures, cost-benefit considerations, and ensuring inclusivity in reaching the most vulnerable groups. Timing is also a critical factor—early interventions can prevent long-term consequences and reduce costs, but they require careful targeting of high-risk individuals. Balancing resource allocation with the needs of those most at risk is essential for effective pro-active and unified policies.

9 Conclusion

This study investigates the distribution of shock persistence in the domains of labor and health and its link to the ex-ante probability of facing those shocks. Using extensive administrative data on the entire Dutch population, we estimate the probability that individuals who have experienced a shock will still be dealing with this situation a year later. Our findings indicate that shock persistence is highly predictable, particularly in the labor domain, and displays significant heterogeneity across individuals. We observe a positive correlation between the probability and impact of a shock, revealing a group of individuals that is particularly vulnerable and poorly resilient to adverse events.

By integrating our findings with previous research, we identify a triple-layered accumulation of risk: (1) ex-ante probabilities of shock occurrence are correlated across domains, (2) the probability of a shock and its impact are positively correlated, and (3) the experience of multiple shocks amplifies the likelihood of persistent shock impacts. Together, these layers underscore the multi-layered nature of risk inequality of labor and health setbacks.

This paper leaves several promising avenues for future research. First, a deeper exploration of underlying shock characteristics could yield valuable insights. While this study examines some factors, such as social benefit type, raise of healthcare expenditures, and primary type of care (physical or mental), there likely exists considerable shock heterogeneity. Future research could, for instance, investigate the income loss experienced by the individual upon job loss, as well as the medical specialization for the health shock. Second, examining the trade-off between the quantity of predictive variables and model accuracy presents another fruitful direction. This study leverages over 500 variables, but further analysis is needed to identify the most crucial variables to achieve reasonable levels of predictive power. Finally, more in-depth research into the effectiveness and timing of preventive measures for individuals with the highest persistence scores could inform targeted policy interventions.

References

Acevedo, P., Mora-Urda, A. I., and Montero, P. (2020). Social inequalities in health: duration of unemployment unevenly effects on the health of men and women. *European Journal of Public Health*, 30(2):305–310.

Acheampong, A. O. and Opoku, E. E. O. (2024). Analyzing the health implications of rising income inequality: What does the data say? *Economics of Transition and Institutional Change*, 32(4):1003–1035.

Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages. *The Economic Journal*, 111(475):F585–F606.

Athey, S., Simon, L. K., Skans, O. N., and Yakymovych, Y. (2024). The heterogeneous earnings impact of job loss across workers, establishments, and markets. *The Institute for Evaluation of Labour Market and Education Policy*, 2024(10).

Atkinson, A. B. and Bourguignon, F., editors (2000). *Handbook of income distribution*, volume 1. Elsevier.

Atkinson, A. B. and Bourguignon, F. (2014). *Handbook of income distribution*, volume 2A-2B. Elsevier.

Bentolila, S., Dolado, J. J., and Jimeno, J. F. (2019). Dual Labour Markets Revisited. *IZA Discussion Paper*, 12126.

Cammeraat, E., Hekkelman, B., Kastelein, P., and Vissers, S. (2023). Predictability and (co-)incidence of labor and health shocks. *CPB Discussion Paper*, 453.

Chetty, R. (2009). Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics*, 1(1):451–488.

Danesh, K., Kolstad, J., Parker, W., and Spinnewijn, J. (2024). The Chronic Disease Index: Analyzing Health Inequalities Over the Lifecycle. *NBER Working Paper*, 32577.

García-Gómez, P., Kippersluis, H. V., O'Donnell, O., and Doorslaer, E. V. (2013). Long-Term and Spillover Effects of Health Shocks on Employment and Income. *Journal of Human Resources*, 48(4):873–909.

Gregg, P. and Tominey, E. (2005). The wage scar from male youth unemployment. *Labour Economics*, 12(4):487–509.

Gregory, V., Menzio, G., and Wiczer, D. (2025). The alpha beta gamma of the labor market. *Journal of Monetary Economics*, 150(103695).

Hoskins, S., Johnston, D. W., Kunz, J., Shields, M., and Staub, K. E. (2024). Heterogeneity in the Persistence of Health: Evidence from a Monthly Micro Panel. *IZA Discussion Papers*, 17023.

Kaplan, S. and Garrick, B. J. (1981). On The Quantitative Definition of Risk. *Risk Analysis*, 1(1):11–27.

Kolsrud, J. and Spinnewijn, J. (2024). The Value and Limits of Unemployment Insurance. *LSE Public Policy Review*, 3(2):8.

Kunaschk, M. and Lang, J. (2022). Can Algorithms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic. *IAB-Discussion Paper*, 8/2022.

Mueller, A. and Spinnewijn, J. (2023). The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection. *NBER Working Paper*, 30979.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nilsen, Ø. A. and Reiso, K. H. (2011). Scarring Effects of Unemployment. *IZA Discussion Paper*, 6198.

Saint-Paul, G., editor (1996). *Dual labor markets: a macreoeconomic perspective.* MIT Press, Cambridge, Mass.

Van Den Berg, G. J., Kunaschk, M., Lang, J., Stephan, G., and Uhlendorff, A. (2023). Predicting Re-Employment: Machine Learning Versus Assessments by Unemployed Workers and by Their Caseworkers. *IZA Discussion Paper*, 16426.

A Methodological Details

A.1 Information in Data Set

Table A.1 summarizes the variables included in the data set, as reported in Cammeraat et al. (2023).

Table A.1: Information in data set

Domain	Variables
Demographic and socioeconomic characteristics	Age; gender; marital status; household composition; migration background; home-ownership status; residential location; educational attainment.
Employment, income, and wealth	Employment status, contract type and economic sector; hours worked (contracted and excess); primary source of income; earnings from (self-)employed labor and wealth; fiscal transfers; paid taxes on income and wealth; unemployment, disability, old age and health insurance premiums; transfers to other households; household disposable income and income before tax; household assets aggregated by broad categories (bank account balances, stocks and bonds, real estate, privately owned firms, and miscellaneous assets); household liabilities aggregated by broad categories (mortgage, student, and other debt); indicator for problematic debt (default on mandatory health insurance premium payment).
Healthcare treatments	Healthcare expenditures covered by default healthcare insurance, aggregated by various broad categories (such as hospital care, intensive care, mental health care, general practitioner, pharmaceutics, dental care, birth care, geriatric care, paramedical care, long-term care, in-home care and care abroad); number of Diagnosis Treatment Combinations (DBC, registration unit of healthcare treatments) by broad category; prescribed medications by broad category; primary medical specializations required for treatments.

A.2 Details Shock and Persistence Definitions

This overview, based on Cammeraat et al. (2023), provides additional details on the variables used in the shock and persistence definitions.

Social benefits The social benefits considered in the labor shock and persistence definitions are unemployment, social assistance, illness/disability and other social security benefits.

- *Unemployment benefits*: Upon job loss, the recipient is entitled to benefits for up to 2 years, depending on the duration of their employment history. For the first two months, the benefit amounts to 75% of the monthly wage, and 70% thereafter. In certain labor agreements, this is topped up to 100% by the employer.

- *Social assistance benefits:* One is entitled to social assistance benefits when one's income and wealth are both below some social minimum thresholds. For a single adult between 21 and the statutory retirement age, the income threshold is set at 70% of the minimum wage.

- *Illness benefits:* Employees without a fixed contract or unemployed people who get ill can apply for illness benefits for up to 2 years. In most cases the amount equals 70% of the wage in the year prior to illness.

- *Disability benefits:* Employees who are considered disabled for more than 35% are eligible for disability benefits, which can amount to up to 75% of their previous salary.

- *Other social security benefits:* This includes various other social security benefits, such as benefits for young disabled people, older and partially disabled unemployed people, and older and partially disabled former self-employed persons.

Health expenditures These are the annual healthcare expenses covered by the mandatory basic health insurance for nearly all Dutch residents. These expenses reflect the actual costs reimbursed by health insurers. We exclude expenditures related to general practitioners and childbirth care.

A.3 Machine Learning Algorithm Parameters

Table A.2 shows the values we set for *LightGBM* package parameters. If a parameter is not listed we use the default package setting. Gradient boosting methods are known to be prone to overfitting, which is why many of our parameter choices are aimed at mitigating overfit. Rather than doing an optimized parameter search, we choose our parameters to work well with the size and type of data we use, which means that similar performance can be expected if a different set of individuals would be selected. Still, a slight overestimate on our test set is possible since performance was measured there.

Table A.2: *LightGBM* package parameters

Parameter	Value	Comment
Number of boosting iterations	150	More leads to overfit as errors move to zero.
Shrinkage rate	0.1	This is a commonly used value to make sure learning is not too erratic.
Maximum leaves per tree	40	More leaves allows for more complex variable interactions, but leads to more overfit as well.
Minimum observations per leaf	22/37	Increasing this parameter significantly reduces overfit because too small leaf size allows fitting highly specific cases. This minimum should be proportional to the number of observations in the train set (in our case it is set at $\sim 0.01\%$), left is for labor $t+1$ and right is for health $t+1$.
Early stopping rounds	15	This reduces pointless training by stopping when the validation score does not improve enough after the chosen amount of rounds.
Bagging fraction	0.9	Another common way to reduce overfit by leaving out a random part of the train set each iteration, allowing more data variation.
Feature fraction	0.9	Similar rationale to bagging fraction, this leaves out a random part of the variables each iteration, allowing more variable variation.
Lamba L1/L2 style regularization	0.01	Reduces overfit by reducing leaf weights.

A.4 Machine Learning Data Pipeline

Each data point in our train and test sets corresponds to a unique individual in a given year. The train sets include $6,716,500$ unique individuals, while the test sets include $6,716,500$ different individuals. Applying sample selection and requiring computable shock definitions as described in section 3, leaves us with $12,607,903$ observations. For both shocks, we select observations where the shock occurred in year t and persistence is computable in year $t+1$, resulting in $212,477$ and $365,329$ observations in the test sets for labor and health shock persistence, respectively. The same procedure results in similar numbers for the train sets.

Each individual has both time-invariant and time-dependent variables. For time-dependent variables, we include the values for the shock year and three lags, resulting in 1,681 variables per observation when combined with the time-invariant variables. The variables can include categorical and missing values, both of which are conveniently handled by the *LightGBM* package. The second and third lags of time-dependent variables include many missing values because those lie outside our dataset for observations in the first two years. To avoid data leakage, no time-dependent information from the prediction year is included. Although an individual may have multiple observations across years, our method ensures that all their data is either in the train or test set, preventing data leakage. Year fixed effects may be captured by the year variable, though we do not observe any.

For both shocks, we train a single model on the respective train set to make predictions for all observations in the respective test set. We thus end up with $212,477$ observations of individuals that suffered the labor shock in year t for which we have both the persistence realization (binary value) and persistence probability estimate (continuous score) in year $t+1$. Similarly for the $365,329$ observations in the health shock test set.

A.5 Variable Importance

While we cannot infer causal relationships between input variables and outcomes predicted by the machine learning methods that we use, we can report variable importance. Variable importance is an opaque way of expressing relative significance of variables for producing predictions. It is opaque because often importance is wrongly attributed to categorical variables or when variables are correlated with each other. A variable importance ranking should therefore not be taken at face value.

Table A.3 lists the top 5 variables with the highest variable importance for the prediction of either labor or health shock persistence. We see these are dominated by variables with direct relation to the shock in the same domain. The first cross-domain variables that we see in the ranking are pharmacy expenses as 19th most important for labor shock persistence and sector collective labor agreement as 12th most important for health shock persistence. Socioeconomic category is only 26th most important for health shock persistence prediction.

Table A.3: Top variable importances of machine learning models

Importance	Labor shock persistence	Health shock persistence
1^{st}	Socioeconomic category	Specialisms with highest expenses
2^{nd}	Working hours (part-time factor)	Pharmacy expenses (covered)
3^{rd}	Year of birth	Work disability duration
4^{th}	Working hours category	Number of clinical admissions
5^{th}	Sector collective labor agreement	Hospital expenses (covered)

A.6 Additional Performance Evaluation

We repeat the performance evaluation from figure 2 for sub-samples of different person characteristics. Table A.4 shows that the regression of persistence realizations on persistence probability estimates yields similar results across the categories and characteristics for both shocks.

Figure A.1 depicts the regression results as well as the percentile bins for the sub-sample of individuals who had either a relatively low or a relatively high predicted ex-ante probability of receiving a shock. Since the prediction performance is comparable, we conclude that the persistence of shocks can be predicted equally well regardless of whether someone received a shock against the odds or whether someone received a shock and was likely to receive it. However, we do observe that the persistence probabilities for the group with a low ex-ante shock probability (panels a and c) are less concentrated in the extremes of the probability space compared to those with a high ex-ante shock probability (panels b and d).

In figure A.2 we assess whether the prediction performance differs between the type of social benefits or healthcare that someone receives. While the results are broadly similar, we do observe some differences. Firstly, the persistence probabilities are much higher for people who receive benefits related to disability insurance or social assistance (panel b and c) compared to unemployment insurance (panel a). Furthermore, the R-squared for the group that receives social assistance is considerably lower compared to the two other main types of social benefits. Secondly, the persistence probabilities of people who receive mental healthcare are condensed to the middle of the probability space while those receiving physical healthcare are much more spread out. The R-squared is also considerably lower, which suggests that the prediction performance might be worse for mental healthcare.

Next, we collapse the continuous persistence probabilities to a binary outcome by choosing, for each persistence definition, a probability threshold above which we predict a persistence realization and below which we predict no persistence realization. Table A.5 presents various performance metrics that are often used when evaluating the prediction quality in

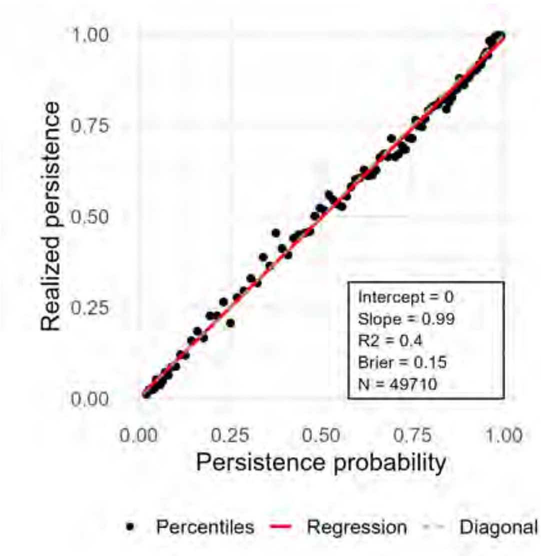
Table A.4: Performance evaluation metrics for sub-samples of different person characteristics observed in the year before the shock year.

	Labor shock				Health shock			
	Intercept	Slope	R2	Brier	Intercept	Slope	R2	Brier
Gender								
Male	0.01	0.98	0.43	0.14	-0.01	1.01	0.15	0.21
Female	0.00	0.99	0.44	0.13	-0.02	1.05	0.19	0.20
Birth cohort								
1953 - 1957	0.02	0.99	0.41	0.07	0.01	1.00	0.15	0.21
1958 - 1962	0.01	0.99	0.40	0.11	-0.01	1.02	0.15	0.21
1963 - 1967	0.01	0.98	0.38	0.13	-0.01	1.03	0.16	0.21
1968 - 1972	0.01	0.99	0.39	0.14	-0.02	1.05	0.16	0.21
1973 - 1977	0.01	0.97	0.38	0.15	-0.02	1.04	0.18	0.20
1978 - 1982	0.00	1.00	0.42	0.15	-0.02	1.05	0.21	0.19
1983 - 1987	0.01	0.99	0.41	0.15	-0.02	1.06	0.21	0.19
1988 - 1992	-0.01	0.99	0.40	0.15	-0.01	1.04	0.19	0.20
Education level								
High	0.00	0.99	0.42	0.14	-0.02	1.05	0.20	0.19
Middle	0.01	0.98	0.42	0.14	-0.02	1.04	0.17	0.20
Low	0.00	0.99	0.42	0.12	0.00	1.01	0.15	0.21
Unknown	-0.01	1.01	0.42	0.09	-0.01	1.03	0.17	0.20
Household composition								
Couple with children	0.00	0.99	0.42	0.14	-0.01	1.04	0.17	0.20
Couple without children	0.00	1.00	0.48	0.12	-0.01	1.04	0.18	0.20
Single with children	0.00	0.98	0.40	0.13	-0.01	1.02	0.15	0.21
Single without children	0.01	0.99	0.44	0.13	-0.01	1.03	0.16	0.21
Migration background								
Dutch origins	0.00	0.99	0.45	0.13	-0.01	1.03	0.17	0.20
Child of migrant inside EU	0.01	0.98	0.39	0.14	-0.01	1.02	0.17	0.20
Child of migrant outside EU	0.00	0.99	0.41	0.12	-0.02	1.04	0.19	0.20
Migrant inside EU	-0.02	1.02	0.46	0.13	0.00	1.01	0.15	0.21
Migrant outside EU	0.02	0.98	0.39	0.15	-0.01	1.05	0.19	0.20
Home ownership								
Own house	0.00	0.99	0.45	0.13	-0.01	1.04	0.17	0.20
Rental house	0.00	0.99	0.45	0.13	-0.01	1.03	0.16	0.21
Rental house with rent allowance	0.02	0.98	0.38	0.12	-0.01	1.02	0.15	0.21
Wealth quintile								
First	0.01	0.99	0.43	0.14	-0.01	1.04	0.17	0.20
Second	0.01	0.99	0.42	0.13	-0.01	1.03	0.16	0.21
Third	0.01	0.99	0.45	0.13	-0.02	1.04	0.18	0.20
Fourth	0.00	0.99	0.45	0.13	-0.01	1.03	0.16	0.20
Fifth	0.00	0.99	0.43	0.13	-0.01	1.03	0.16	0.20

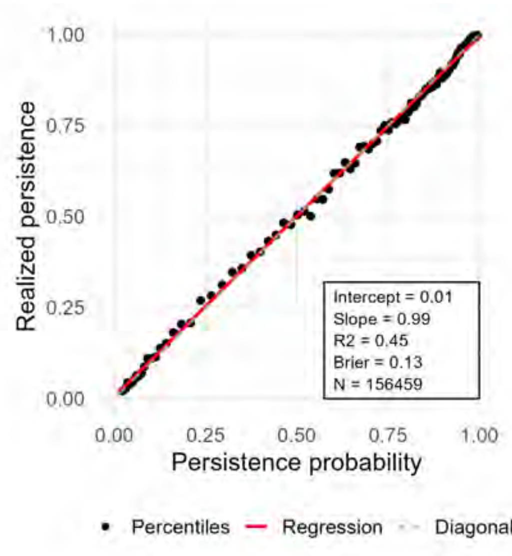
Figure A.1: Regression of persistence realizations on persistence probability estimates for the labor and health shock, split up by the level of the ex-ante shock probability estimate. Low ex-ante probability corresponds to the sample of individuals with an ex-ante shock probability estimate less than two times the population average, while high ex-ante probability corresponds to the group of individuals with an ex-ante shock probability estimate that is more than two times the population average.

Labor shock

(a) Low ex-ante probability

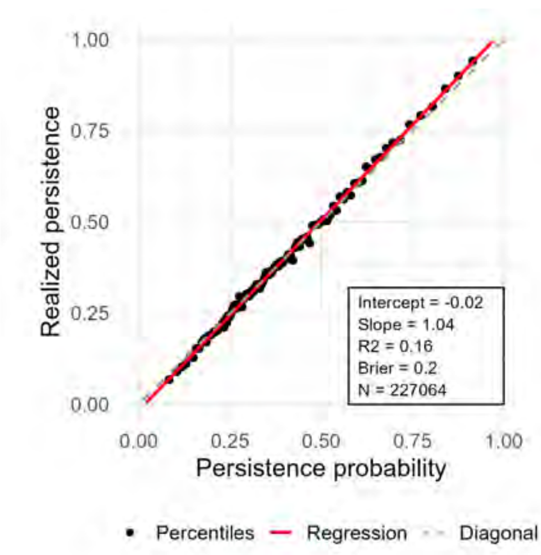


(b) High ex-ante probability



Health shock

(c) Low ex-ante probability



(d) High ex-ante probability

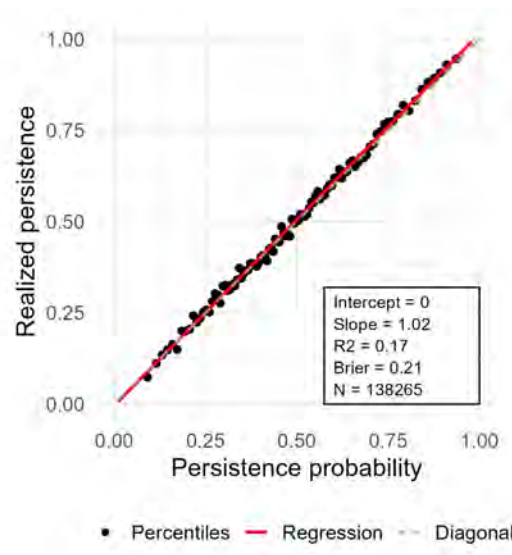
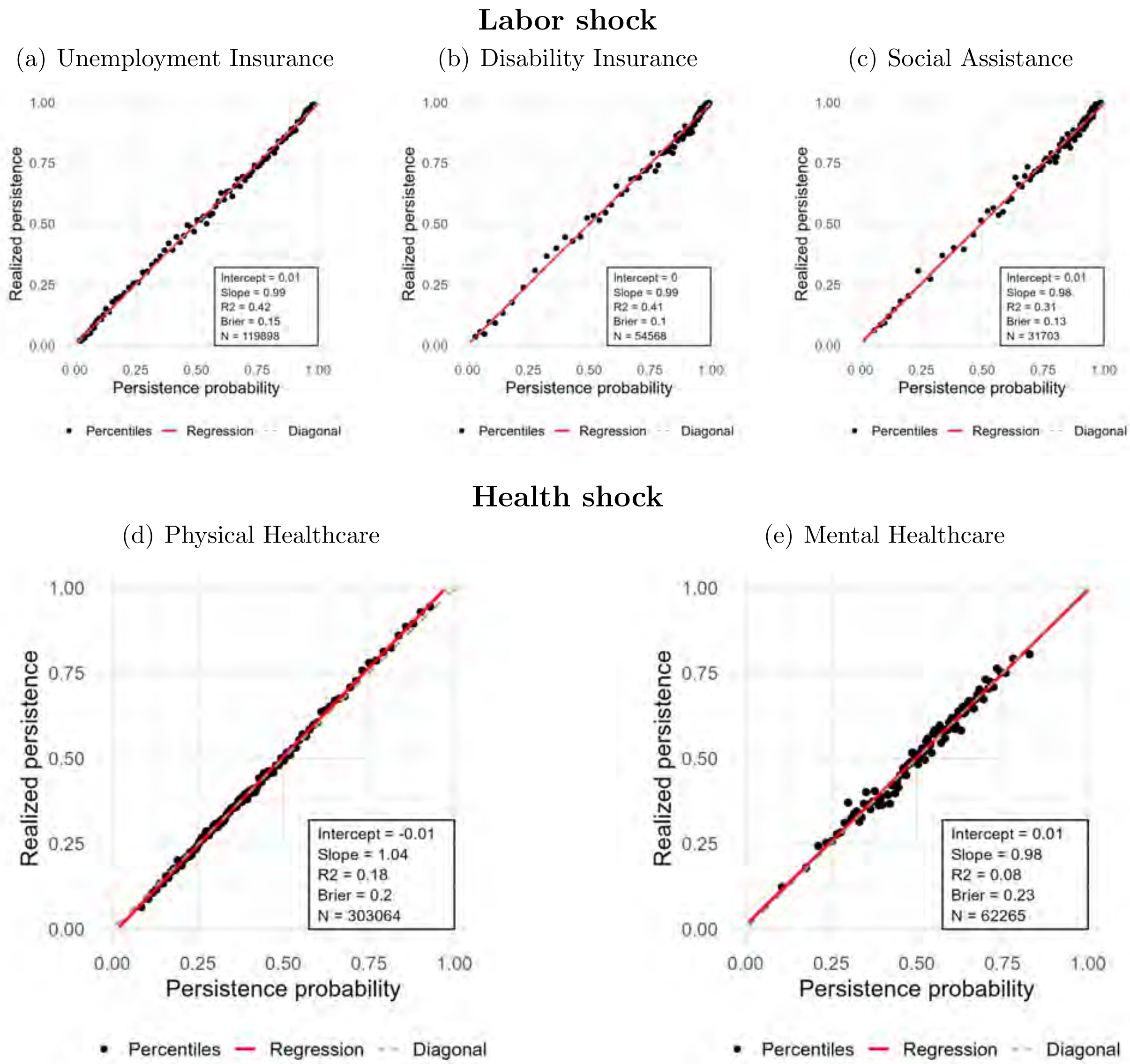


Figure A.2: Regression of persistence realizations on persistence probability estimates for the labor and health shock, split up by the type of shock realization.



classification exercises. The AUC (Area Under the Curve) measures the area under the ROC (Receiver Operating Characteristic), which plots the true positive rate against the false positive rate across thresholds. It ranges from 0.5 (no predictive power) to 1 (perfect prediction). The other metrics are calculated at the threshold maximizing the F1-score on the test set. The F1-score is the harmonic mean of Precision (true positives/total positives) and Recall (true positives/actual positives), while Accuracy reflects the proportion of correct predictions.

Table A.5 shows that both persistence definitions achieve good values for the AUC metric, indicating that both prediction models have adequate discerning capabilities. The prediction performance is better for the labor shock than for the health shock, a result similar to shock prediction models in Cammeraat et al. (2023).

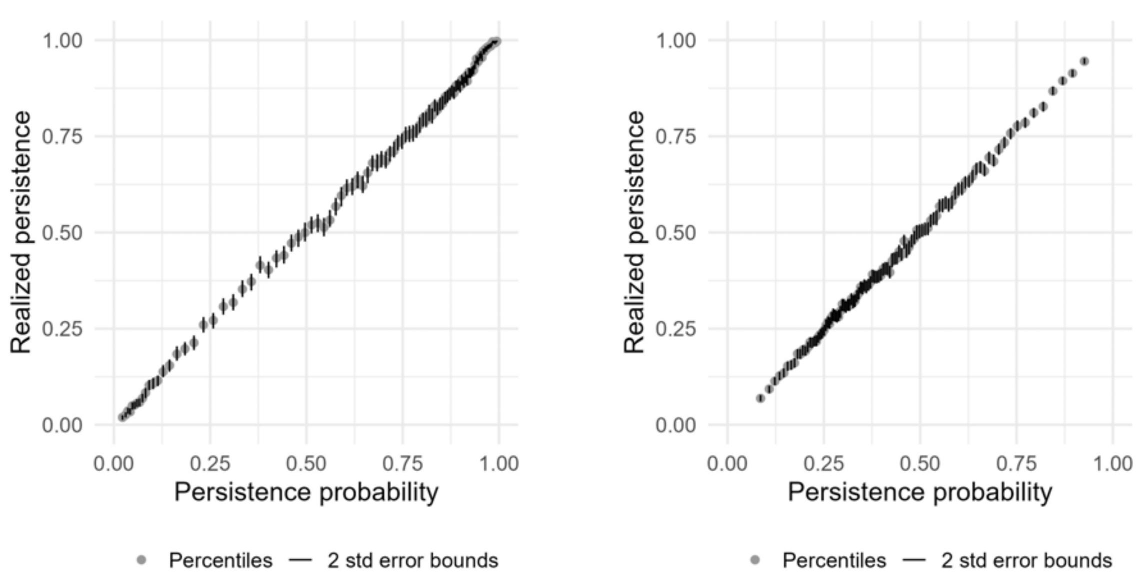
Table A.5: Classification performance metrics for both shocks.

Shock	AUC	F1-score	Precision	Recall	Accuracy
Labor	0.88	0.86	0.80	0.93	0.81
Health	0.74	0.65	0.54	0.82	0.62

Finally, we investigate to what extent random variation in the sample affects the models' probability estimates. Figure A.3 shows, for both shocks, the variation in realization prevalence per percentile bin for 1,000 bootstrap samples. As we see the pattern around the 45° line remains strong, implying a robustness to sample variation.

Figure A.3: Variation of probability estimate and realization prevalence per percentile bin for 1,000 bootstrap samples of both shocks.

(a) Labor shock (b) Health shock



B Supplementary Analysis

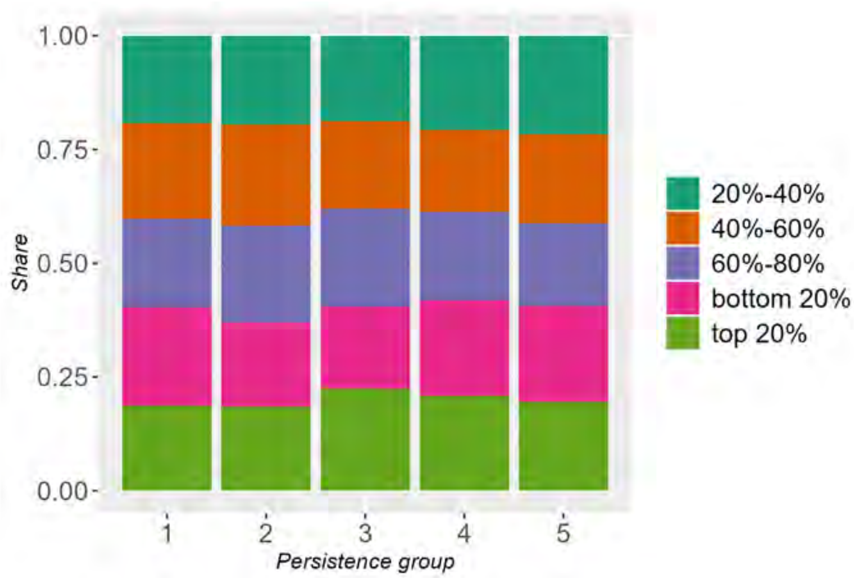
B.1 Persistence Distribution Conditional on Shock Type

Section 6 presents the distributions of shock persistence probabilities for the labor and health shock. However, this overlooks potential heterogeneity in the shock type. Therefore, in this section, we split up the distributions so that specific groups can be highlighted.

B.1.1 Healthcare Shock Magnitude

The healthcare shock is defined as crossing a healthcare expenditures threshold, but the amount by which the threshold was exceeded will differ among individuals. Figure B.1 shows the shares of magnitude quintiles of the initial healthcare expenditures shock for each quintile of predicted persistence (the lowest persistence scores on the left). It shows that the magnitude of the health expenditures shock does not show a strong relation with the persistence probability.

Figure B.1: Magnitude of healthcare expenditures shock in year t across persistence groups.



B.1.2 Probability Distributions Conditional on Shock Type

The individuals for whom we predict shock persistence experienced different types of shocks. These underlying shock types are not visible in the density plots of figure 3. To offer a more detailed perspective, we distinguish between the types of social benefits in figure B.2, and between physical and mental healthcare expenditures for the health expenditure shock in figure B.3.

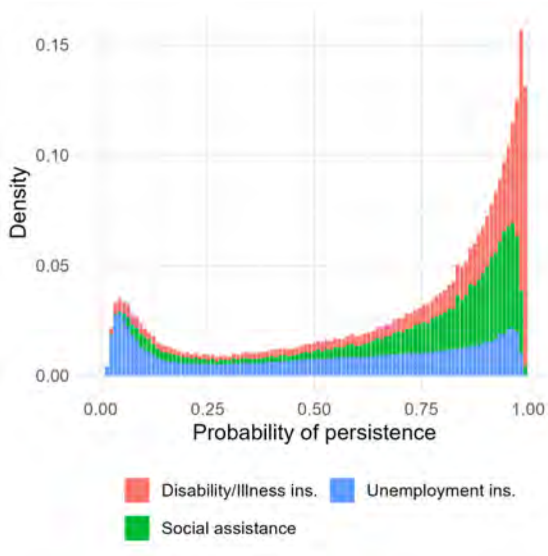
Figure B.2(a) displays the persistency density, broken down by the three underlying shock types: disability insurance, unemployment insurance, and social assistance benefits. Figures B.2(b) to B.2(d) presents these densities separately. A clear bimodal distribution is observed for individuals dependent on unemployment benefits, although a significant group also appears in the middle range. In contrast, the persistence distributions for disability insurance and social assistance are more right-skewed, reflecting the nature of these benefits.

Figure B.3(a) shows the density of persistence for the health shock, distinguishing between physical and mental health shocks. This categorization is based on whether the majority of the increase in expenditures is attributed to one category.[12] Figures B.3(b) and B.3(c) present the densities separately. These plots indicate that, on average, the persistence of physical health shocks is lower than that of mental health shocks. However, the observations with the highest persistence scores are associated with physical health shocks.

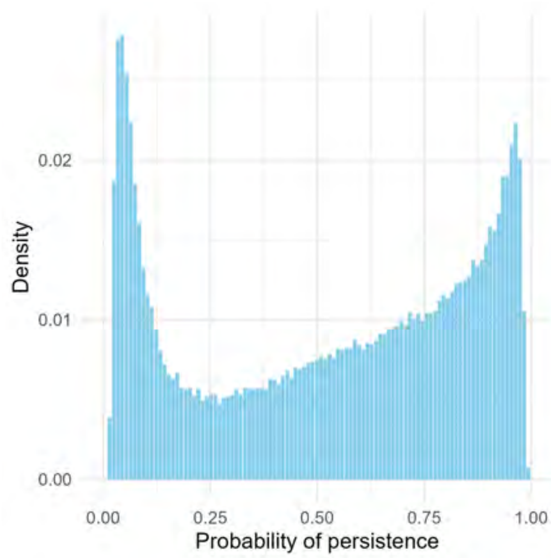
[12] A robustness check revealed that only a small fraction of cases involve mixed shocks within a 40-60 bandwidth.

Figure B.2: The histograms display the density of persistence probabilities for the labor shock (a) and the three underlying types of benefits (b,c,d).

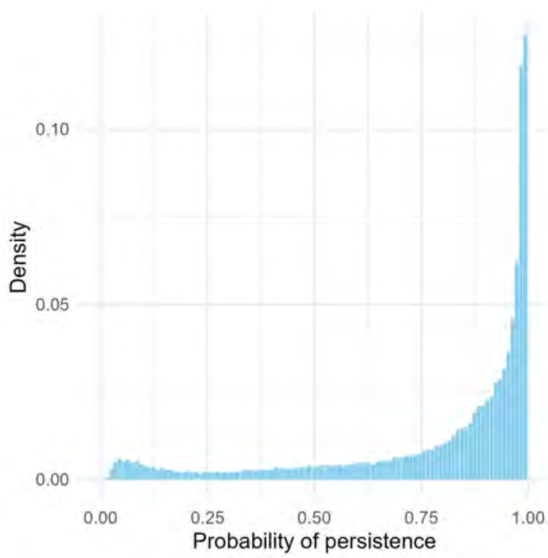
(a) Total



(b) Unemployment Insurance



(c) Disability Insurance



(d) Social Assistance

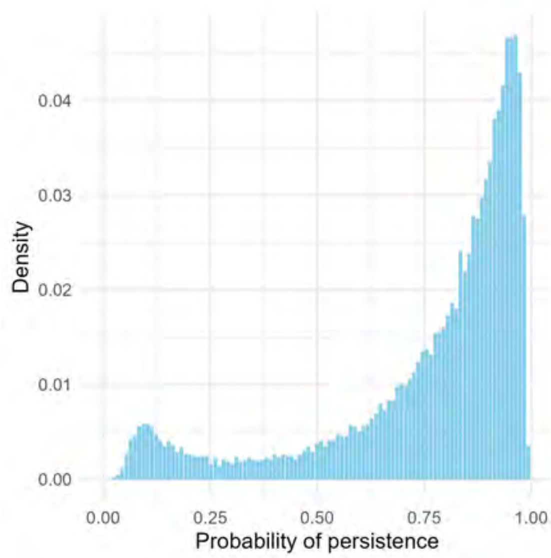
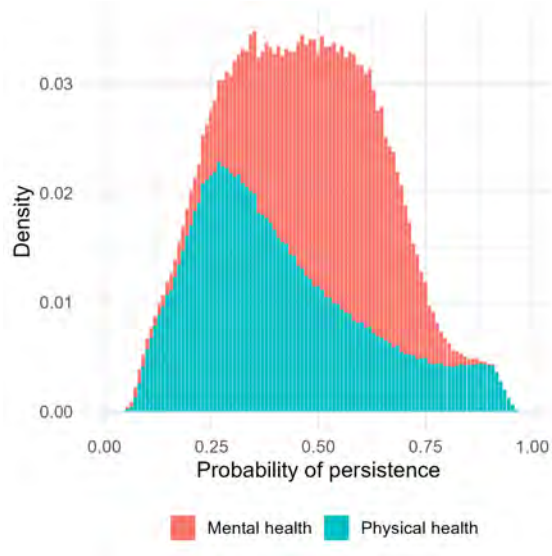
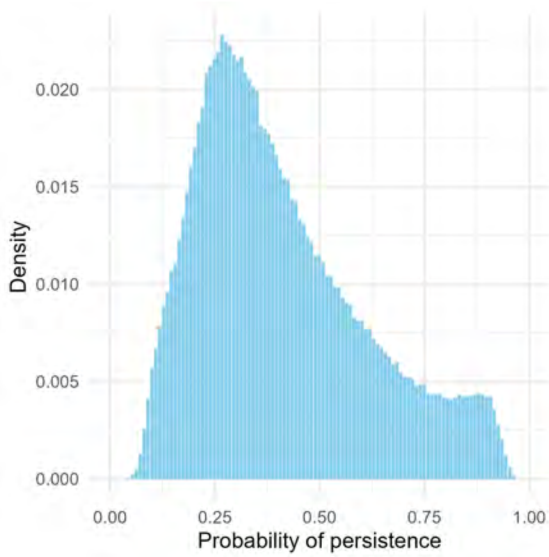


Figure B.3: The histograms display the density of persistence probabilities for the health shock (a) and conditional on the type of healthcare, physical or mental (b,c).

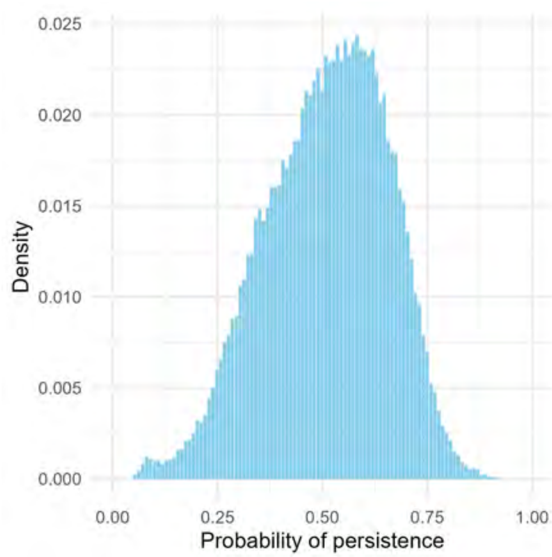
(a) Total



(b) Physical Health Shock



(c) Mental Health Shock

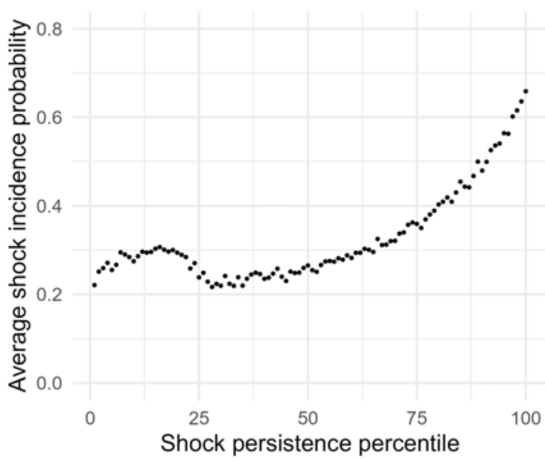


B.1.3 Shock Exposure and Shock Persistence Conditional on Shock Type

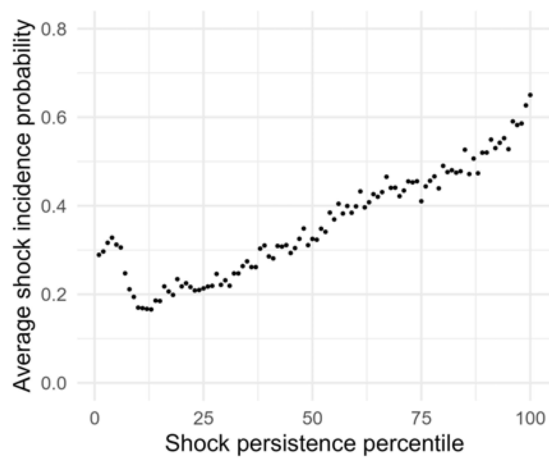
Figure B.4 and figure B.5 show the relationship between ex-post persistence probabilities and ex-ante shock incidence probabilities as shown in figure 4 for the different underlying shock types. The positive relationship between ex-ante probability and ex-post persistence is visible for both unemployment and disability insurance, although there exists a group of people with a high shock probability and a low persistence probability (in the left of the distribution). For social assistance the relation is flat. It is unclear whether this is due to relatively poor performance of our prediction model for this benefits type (see the discussion in appendix A.6) or whether it is indeed the case that the ex-post probability of persistence is orthogonal to the ex-ante probability of shock incidence.

Figure B.4: The figures show the relationship between ex-post persistence probabilities and ex-ante shock incidence probabilities conditional on the different types of social benefits. The x-axis represents percentiles of estimated persistence probabilities, while the y-axis shows the average shock incidence probability for individuals in each percentile.

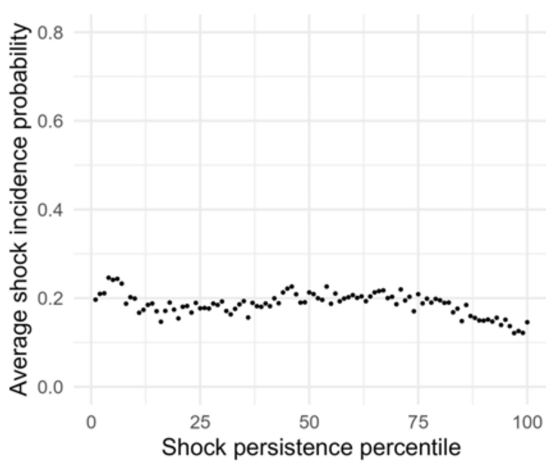
(a) Unemployment insurance



(b) Disability insurance



(c) Social assistance

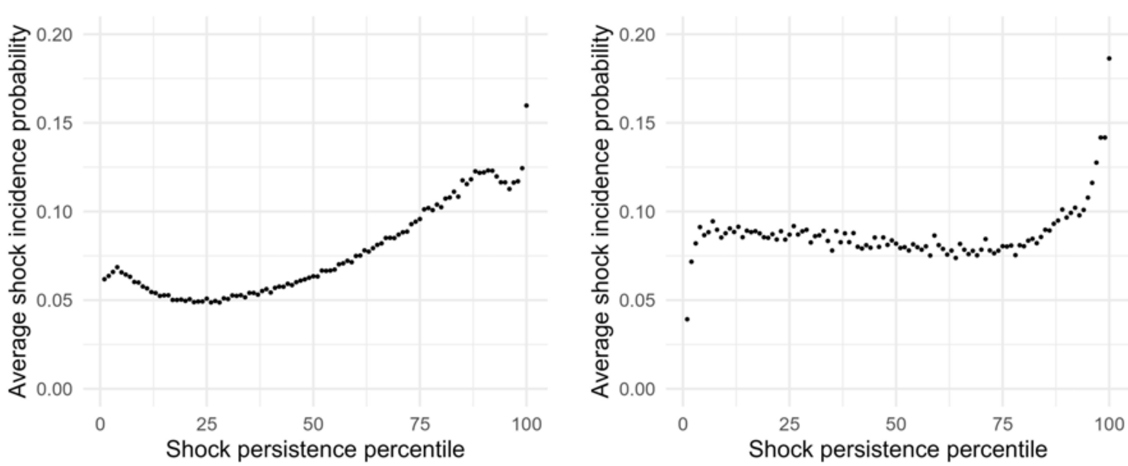


The correlation between persistence and incidence probabilities is positive for both types of healthcare, but there are striking differences across the distribution. Similar to the labor shock, there exists a group of people receiving physical healthcare with a high shock probability and a low persistence probability (in the left of the distribution). The relationship between the ex-ante and ex-post probabilities is mostly monotonic outside of this region. For mental healthcare the ex-ante and ex-post probabilities appear orthogonal to each other for the majority of the persistence probability space, with an exception at both extremes where the relationship is positive.

Figure B.5: The figures show the relationship between ex-post persistence probabilities and ex-ante shock incidence probabilities conditional on the different types of healthcare. The x-axis represents percentiles of estimated persistence probabilities, while the y-axis shows the average shock incidence probability for individuals in each percentile.

(a) Physical health shock

(b) Mental health shock



B.2 Persistence Two Years after Shock

We alter the persistence definitions to two years. That is, in the labor domain the persistence definition becomes that social benefits remain the individual's primary source of income two years after the initial event. We do not impose a condition on the first year after the shock, hence, it is possible that the individual was back at work for a short period.

In the domain of health, the definition becomes that the healthcare costs two years after the shock have not decrease by more than 80% compared to the increase in the shock year. Again, we do not impose a condition on the intermediate year.

Figure B.6 displays the regression of the realized persistence percentiles on the probability estimates for the extended time window. The results show that also in case of the extended time window, the regression line follows the diagonal closely, indicating strong performance. Compared to the base line in figure 2, most performance metrics have decreased slightly. Also, the classification metrics are slightly lower compared to the baseline case, see table B.1.

Figure B.6: Regression of persistence realizations on persistence probability estimates for $t + 2$.

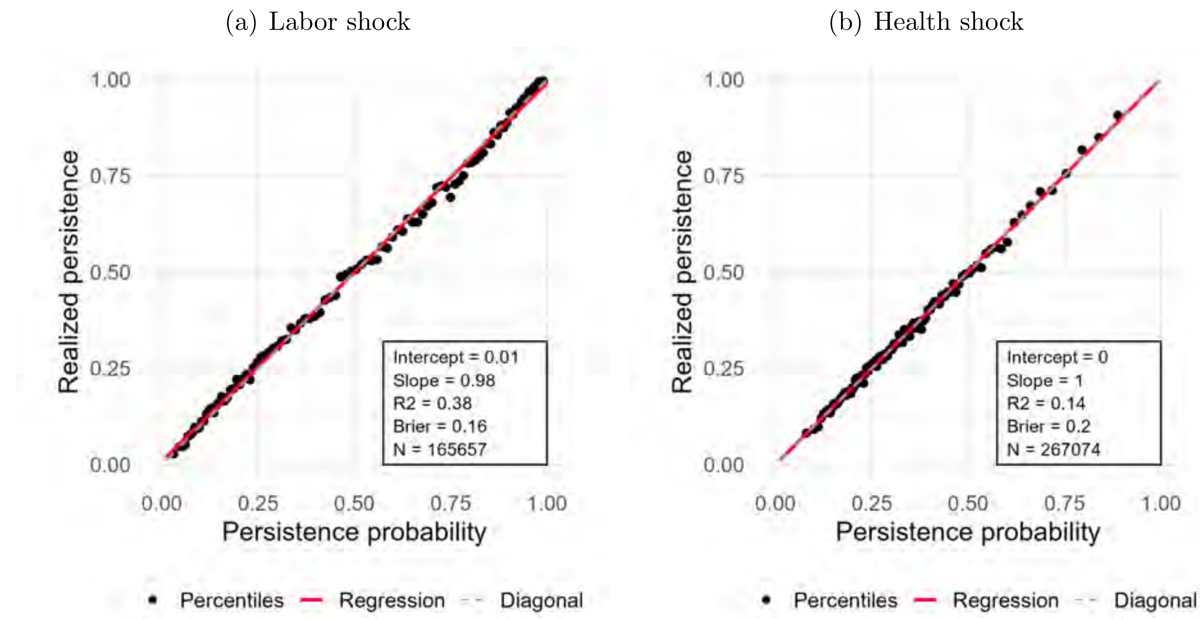


Figure B.7: Variation of probability estimate and realization prevalence per percentile bin for 1,000 bootstrap samples of both shocks for $t+2$.

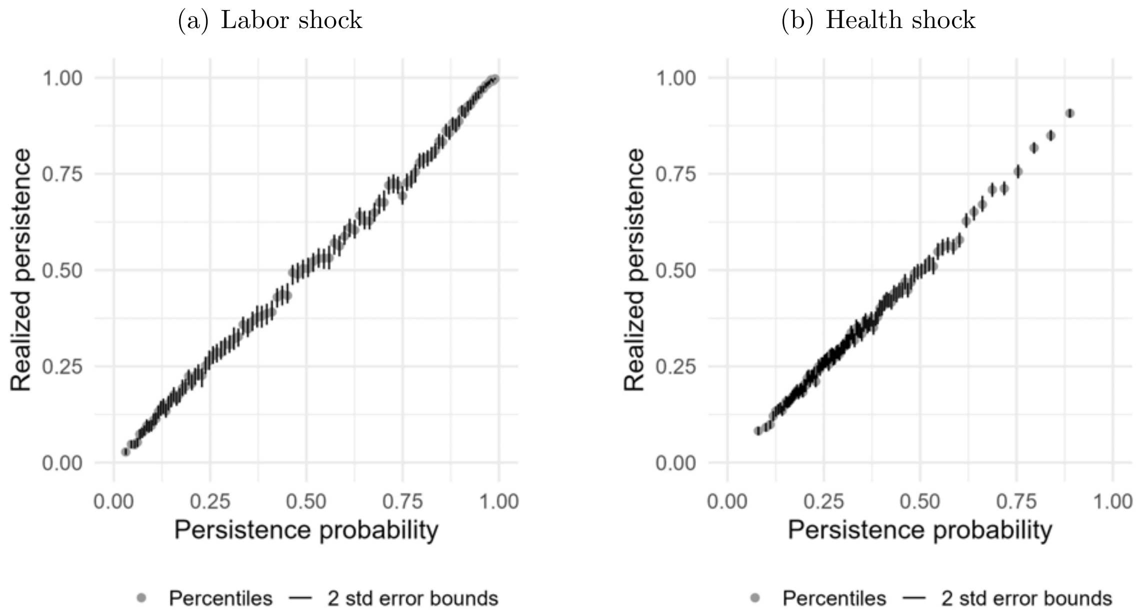


Table B.1: Classification performance metrics for $t+2$.

Shock	AUC	F1-score	Precision	Recall	Accuracy
Labor	0.85	0.78	0.74	0.83	0.76
Health	0.71	0.57	0.46	0.75	0.61

Figure B.8 shows the distribution of the persistence probabilities for the extended time window, contrasted to the original distribution of $t+1$ from figure 3. Generally speaking, we see in both graphs that mass moves to the left, indicating that the probability of shock persistence at decreases over time. For labor, this could be partly explained by unemployment rights running out, incentivizing people to look for work.

Figure B.8: The histograms display the density of persistence probabilities for the labor shock (left panel) and the health shock (right panel) for both $t+1$ and $t+2$.

(a) Labor shock (b) Health shock

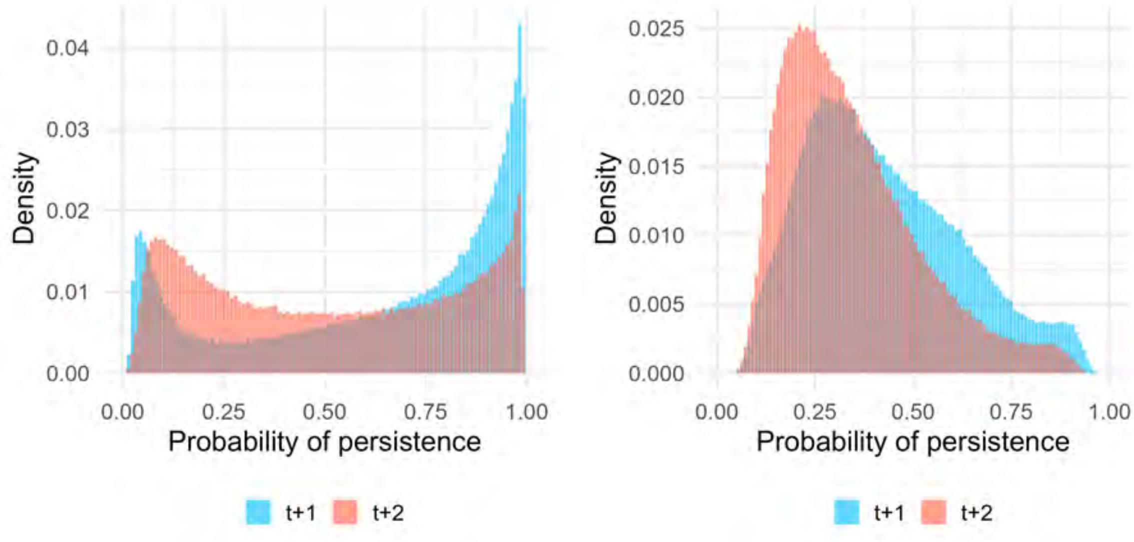
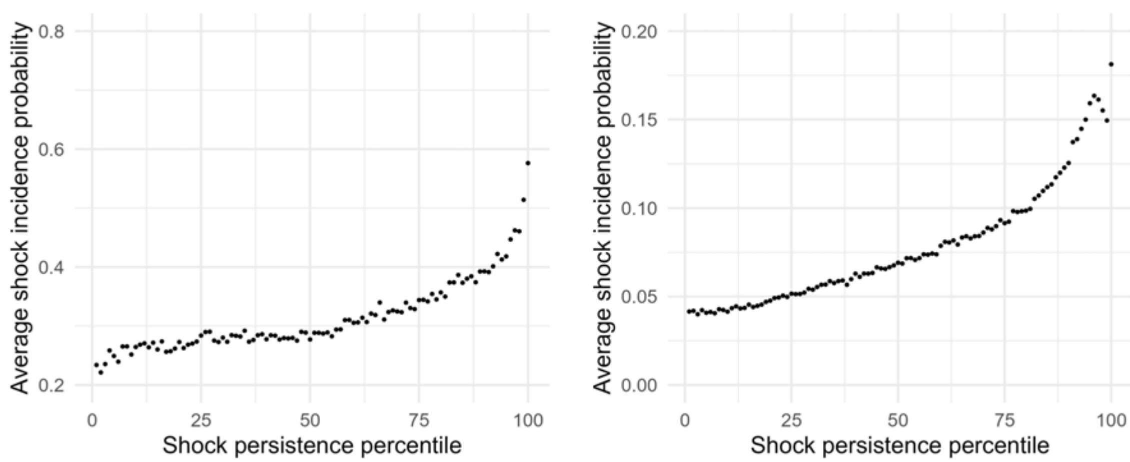


Figure B.9: The figures show the relationship between ex-post persistence probabilities and ex-ante shock incidence probabilities for the extended time window $t+2$. The x-axis represents percentiles of estimated persistence probabilities, while the y-axis shows the average shock incidence probability for individuals in each percentile.

(a) Labor shock (b) Health shock



B.3 Multiclass Analysis

Thus far, our analysis has employed binary definitions for shock persistence, classifying outcomes simply as either persistent or non-persistent. In this section, we examine the trajectories that emerge in the year following the initial shock. For the labor shock, we differentiate between distinct categories of social benefits received: unemployment benefits, disability or illness benefits, social assistance, and other forms of support. For health shocks, we categorize outcomes by the magnitude of reduction in healthcare expenditures. Instead of predicting a binary persistence outcome, we assign a probability distribution across possible states in the post-shock year. Consequently, each individual-year observation is represented by a probability vector that sums to 1, providing a comprehensive view of potential outcomes and their associated likelihoods.

B.3.1 Multiclass Analysis of Labor Shock

Table B.2 shows the realized paths for different types of the labor shock. In general, we do not observe many switches to different types of benefits. Rather, we see that most people who had a unemployment insurance shock find work in the next year, a somewhat smaller amount keeps receiving unemployment benefits, while a small fraction moves to disability insurance or social assistance. Both for disability insurance and social assistance, we observe a much stronger persisting effect.

Figure B.10 presents the predictions from the multiclass analysis for labor shocks. Here, percentiles are constructed based on the probability of shock persistence, resulting in a decreasing probability of returning to work by construction. Figure B.10(a) displays the distribution across all types of labor shocks. One could discern three groups. On the left side of the distribution are the individuals who are very likely to find work again next year. In the middle group are the individuals who have still a reasonable change of finding work, but who are also potentially in some form of social benefits in the next year. The third group on the right end of the distribution are the individuals who are almost certain to rely

Table B.2: Status at $t+1$ for different types of shocks at t, including unemployment insurance (UI), disability/illness insurance (DI), social assistance (SA). The first column presents the distribution of shock types in shock year t as a percentage of the total sample. The remaining columns present the percentage distribution of individuals receiving each type of social benefit in year $t+1$, relative to all individuals who experienced that specific shock type in t.

		Status at $t+1$				
		UI	**DI**	**SA**	**Other**	**Work**
Shock at t						
UI	59.9%	42.0%	6.5%	2.5%	1.3%	47.0%
DI	24.6%	10.0%	65.8%	1.6%	1.6%	21.1%
SA	15.5%	0.8%	0.9%	72.7%	2.4%	23.2%

on disability insurance in the next year.

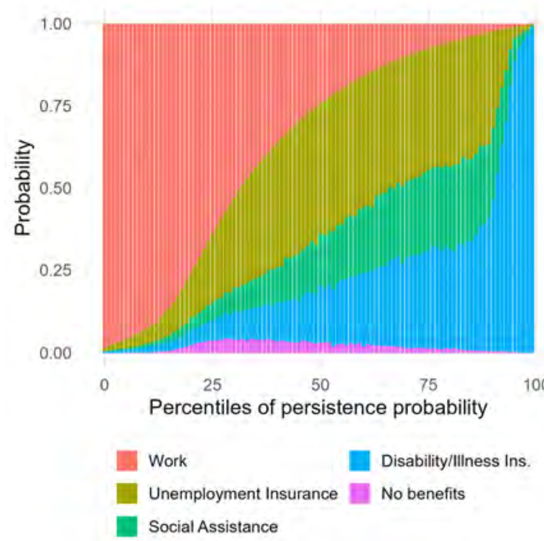
Figure B.10(b) conditions on the individuals who began receiving social assistance benefits. In line with the nature of social assistance benefits, these individuals have a near-zero probability of receiving unemployment and disability benefits in the next year. Of the 50% individuals with the highest persistence probability, the probability of remaining on social assistance exceeds 80%; for the top 25%, this probability rises above 90%.

Figure B.10(c) examines individuals who started receiving unemployment benefits, revealing generally more favorable prospects for transitioning back to work. At the higher end of the persistence distribution, however, we observe an increased probability of transitioning to disability insurance.

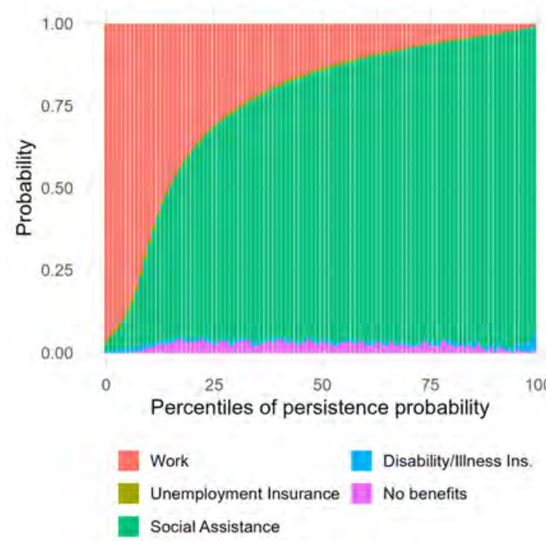
Finally, figure B.10(d) shows outcomes for individuals who began receiving disability insurance. Similar to social assistance, the probability of returning to work is generally lower for this group, although there is substantial heterogeneity. Individuals in the leftmost part of the distribution show a high probability of returning to work within the next year. A middle group has a higher likelihood of remaining on disability insurance but also faces a significant chance of moving to unemployment benefits. At the right end of the distribution, individuals are almost certain to stay on disability insurance.

Figure B.10: Multiclass predictions for labor shock. The figures show the distribution of probabilities assigned to different states in the year after the shock. The underlying observations are sorted into percentiles on the basis of the persistence probability. Panel (a) shows the distribution of the states for all labor shocks, panel (b) for the social assistance shocks, panel (c) for the unemployment insurance shock, and (d) for disability/illness shock.

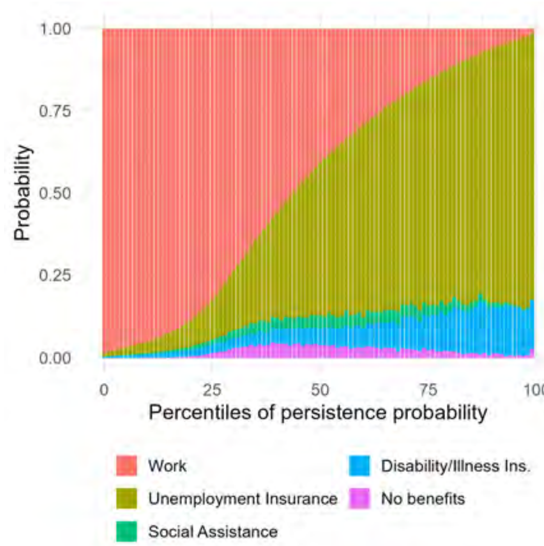
(a) Total



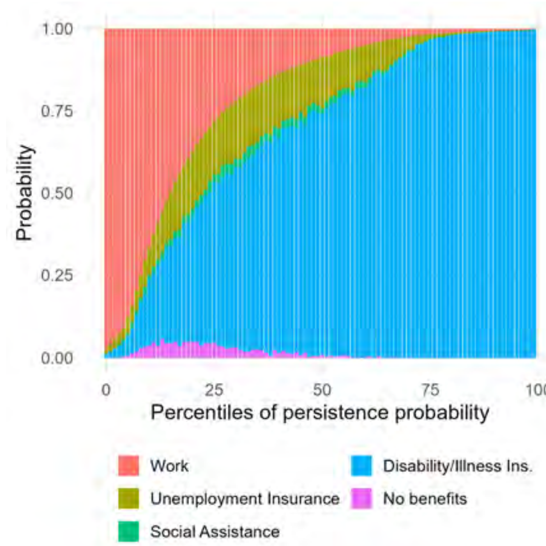
(b) Social Assistance



(c) Unemployment



(d) Disability/Illness



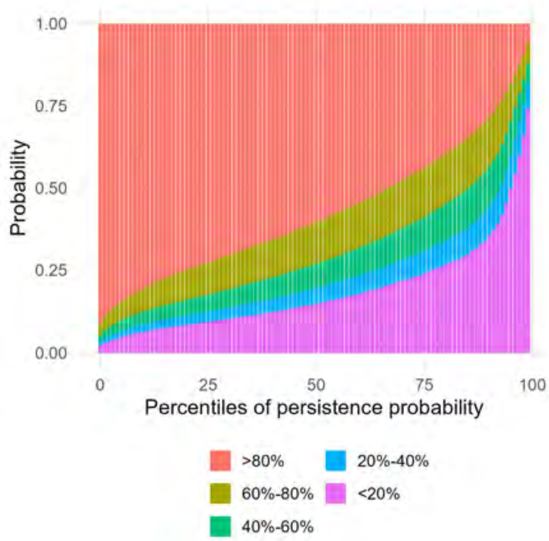
B.3.2 Multiclass Analysis of Health Shock

Figure B.11 presents the predictions from the multiclass analysis for health shocks. Again, percentiles are constructed based on the probability of shock persistence, resulting in a decreasing probability of healthcare expenditure reductions exceeding 80%. Figure B.11(a) displays the distribution across both types of health shocks. Across the board we see that healthcare expenditure reductions exceeding 80% constitute a significant portion. Only on the right of the distribution do we see a group of people with a sharp uptick of healthcare expenditure reductions lower than 20%. Healthcare expenditure reductions between 20% and 80% play only a limited role, and taper off for those at the two extremes of the distribution.

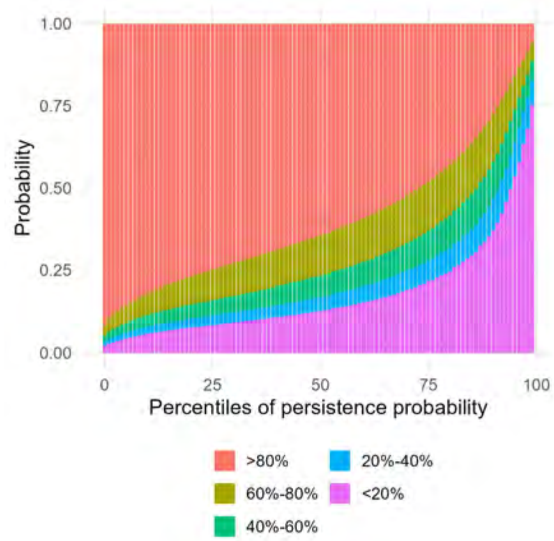
Figure B.11(c) conditions on the individuals who had predominantly mental health expenditures. We see that the worst persistence, with a healthcare expenditure reduction no more than 20%, is a more significant portion across the entire distribution. Figure B.11(b) conditions on the individuals who had predominantly physical health expenditures, which follows a pattern more closely matching the overall distribution.

Figure B.11: Multiclass predictions for health shock. The figures show the distribution of probabilities assigned to different levels of decrease in healthcare expenditures in the post-shock year. The underlying observations are sorted into percentiles on the basis of the persistence probability. Panel (a) shows the distribution for both health shocks, panel (b) for physical health shocks, and panel (c) for mental health shocks.

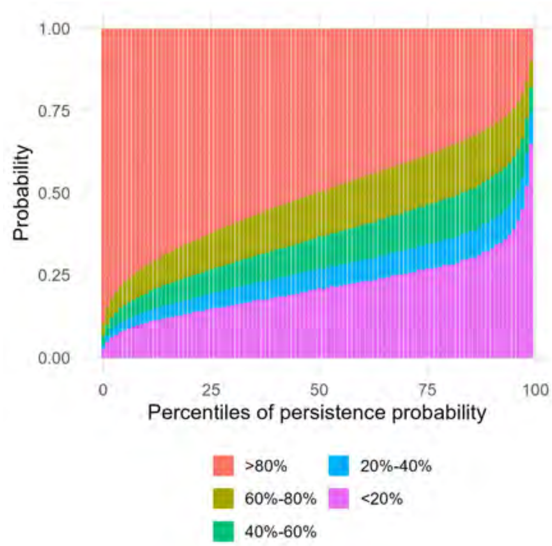
(a) Total



(b) Physical health



(c) Mental health



B.3.3 Performance Evaluation

Table B.3 tabulates the average and maximum absolute prediction errors for the different trajectories after a labor and health shock. The average prediction error is around or below 1%-point, indicating good prediction quality across all trajectories.

Table B.3: Average and maximum absolute prediction errors for different trajectories.

Shock t	**Status** $t+1$	**Mean Absolute Prediction Error (%-pt)**	**Maximum Absolute Prediction Error (%-pt)**
Labor	Work	1.6	6.4
	Unemployment Insurance	0.9	4.1
	Social Assistance	0.5	2.1
	Illness/Disability	0.5	1.5
	No Benefits	0.4	2.4
Health	>80% decrease	0.9	2.9
	60%-80% decrease	0.6	2.8
	40%-60% decrease	0.4	1.6
	20%-40% decrease	0.5	2.9
	<20% decrease	0.6	3.8