CPB Netherlands Bureau for
Economic Policy Analysis

Predictability and (co-)incidence of labor and health shocks

Using machine learning techniques and anonymous data on millions of Dutch people, this study maps out the entire ex-ante probability distribution of a wide range of labor market and health shocks.
This allows us to separate predictable components of shocks, interpreted as ex-ante risk types, from ex-post random components.

We uncover striking levels of ex-ante risk exposure inequality across the population.
Moreover, labor and health risks appear to be strongly related.
These findings offer perspective for targeted prevention policies that provide proactive support to vulnerable groups.

CPB Discussion Paper

Emile Cammeraat, Brinn Hekkelman,
Pim Kastelein, Suzanne Vissers
December 2023

Doi: https://doi.org/ 10.34932/n1bs-qm11

Predictability and (Co-)Incidence of Labor and Health Shocks

Emile Cammeraat[a], Brinn Hekkelman[a], Pim Kastelein[a,b], and Suzanne Vissers[a]

February 23, 2024

Abstract

This paper employs machine learning techniques to estimate the probability that individuals will be confronted with adverse labor and health events. We separate predictable components of shocks, interpreted as ex-ante risk types, from random components by contrasting ex-post shock incidence with ex-ante shock probabilities. Using rich administrative data on the entire Dutch population, we uncover three results. First, shock incidence is predictable especially in the labor domain and to a lesser degree also in the domain of mental and physical health. Second, risk exposure to shocks is very unevenly distributed, persistent over time and correlated with socioeconomic characteristics such as employment status, educational attainment, income and wealth. Third, bringing together risk estimates for twelve different shocks highlights that risk concurrence is sizable, monotone, non-linear and extended across domains. We discuss how our findings can help improve the targeting of prevention policies.

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

Keywords: risk concurrence, labor shocks, health shocks, machine learning, prediction.

[a]CPB Netherlands Bureau for Economic Policy Analysis

[b]Corresponding author: p.kastelein@cpb.nl

We thank Johannes Spinnewijn, Bas ter Weel 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. Special recognition is extended to Hannah Gelblat for her exceptional research assistance.

1 Introduction

People's lives can be disrupted by unexpected adverse events such as job loss or illness. These shocks do not occur solely in isolation, but can rather cascade through a complex web of interactions. Falling ill may trigger unemployment, while struggling with mounting debts can set off mental health struggles (García-Gómez et al., 2013; Roos et al., 2021; Adda et al., 2009). This paper studies two questions regarding the incidence of such shocks. First, whether they occur randomly or whether they carry an element of predictability that could be acted upon. Second, to what degree the risk of incurring one shock is related to the risk of incurring other shocks. We bring together data on millions of people to explore the predictable element of a wide array of labor and health shocks. This unveils the distributions of the predictable risk factors throughout the population. The risks turn out to be very unevenly distributed and strongly correlated with each other.

The materialization of a shock can be viewed as the culmination of both predictable and random components. When analyzing the incidence of shocks it is thus unclear to what extent they were able to be anticipated and to what extent they were due to bad fortune. We use machine learning techniques to disentangle the two components by estimating the ex-ante probability that someone will face a shock. Following Mueller and Spinnewijn (2023), we interpret the predictable component as an approximation of someone's ex-ante latent risk *type*.[1] Contrasting the ex-post realizations with the ex-ante probabilities shows the relative size of the predictable and random components. Furthermore, while the shock realizations constitute a vector of binary outcomes, the shock probabilities trace the entire risk distribution. Putting together the risk distributions of shocks across different domains provides a novel view of risk concurrence and constitutes our main contribution to the literature.

We collect administrative data encompassing labor, health, socioeconomic and demographic information for the entire Dutch population from 2013-2018, comprising a total of

[1] In this paper, we will use the terms '(ex-ante) risk type' and '(ex-ante) risk' interchangeably.

over five hundred variables observed at yearly frequency for over twenty million people. Based on this data we define twelve shock variables that signify a significant worsening of one's circumstances. The set of labor shocks includes considerable income drops, reliance on social benefits and coping with problematic debts, while the set of health shocks includes incurring physical or mental health care expenditures and receiving treatments. We then train prediction models to estimate the probability that someone will be faced with a shock based on what is known about them in prior years. Machine learning is the right tool for this exercise since it is optimized for prediction performance, handles large amounts of data and discerns the complex interactions that are present in the data (Mullainathan and Spiess, 2017). This allows us to accurately estimate risks and in turn to characterize the relationship between the shocks across the labor and health domains.

The trained prediction models achieve promising accuracy in both classification (predicting whether someone will face a shock) and estimation (predicting the probability that someone will face a shock) exercises according to conventional metrics for assessing prediction performance. This is especially the case in the labor domain and to a lesser degree in the domains of mental and physical health. In other words, the predictable component turns out to be larger for labor shocks compared to health shocks, since the ex-ante labor shock probabilities explain more of the variation in ex-post shock incidence.

The estimated probabilities underscore a strong disparity of risk exposure across the population, scaling in a non-linear way when moving up the risk distribution. As an example, for the event that social benefits become one's main income source, we find that people in the highest percentile face over thirty times the average risk while the majority of people face virtually no risk at all. One's rank in the risk distribution is also highly persistent over time, adding credence to the hypothesis that our risk estimates signify latent risk types or individual fixed effects. However, the extent of the risk dispersion only becomes apparent when one has extensive information about individuals. Sensitivity analysis highlights that relying solely on rudimentary person characteristics achieves very poor prediction performance and

little predicted inequality of risk exposure.

Putting the risk distributions of all twelve shocks together, we find that risk concurrence is sizeable, monotone, non-linear and extended across domains. Individuals with a high estimated risk for one particular shock are also more likely to face other shocks. This not only holds for shocks within the same domain, but also for shocks across domains. As an example, people who are most likely to incur a sizable increase in health expenditures are also up to four times more likely, compared to the population average, to have to suddenly rely on social benefits as their main income source. A novel insight is that the positive correlation between risks appears throughout the entire distribution. Moving up the ranks across the risk distribution of one particular shock implies moving up the ranks of other risk distributions as well, regardless of one's initial position in those distributions. Furthermore, our machine learning models predict higher risks for people who recently experienced other setbacks. Risk estimates are especially heightened for labor shocks after the materialization of health shocks, but the effect vice versa is also substantial.

An overarching theme of our results is that there exists a concentrated group of people which bears the most risk, while the majority faces only scarce amounts of risk. The vulnerable group is predominantly made up of people with temporary employment contracts, those with lower levels of education, income and wealth, individuals originating from outside the Netherlands, residents of rental properties, and singles. This underscores the unequal distribution of risks across the Dutch population within the domains of labor and health.

Policy can influence the shapes of the risk distributions and our results suggest that mitigating the excessive conjunction of risks would be desirable. Thankfully, the finding that risk types are estimable can be of help to policymakers. Social insurance policies play a major role in protecting people against setbacks. However, existing policies tend to fall short by responding re-actively to shocks and by treating them in isolation. If we can accurately predict in advance whether a shock will materialize, it may in some cases be more prudent to focus on prevention beforehand rather than assistance afterwards. This is especially the

case if shocks cascade since proactive policy could break up chains of adverse events. While the predictability of shocks alone is not a sufficient condition to conclude that targeted prevention policies are preferable to the status quo, it is a necessary condition.

This paper proceeds as follows. Section 2 discusses the strands of literature that this paper relates to. Section 3 gives a description of the data used to train and test the machine learning models. Section 4 details the shock definitions in the labor and health domains. Section 5 explains the methodology to obtain the risk estimates of individual shocks and evaluates the prediction quality. Section 6 describes the risk distributions for individual shocks, while section 7 shows how the risk distributions of different shocks relate to each other. Section 8 explores the characteristics of high-risk individuals. Section 9 discusses the policy implications of our results. Section 10 concludes.

2 Related Literature

An extensive literature has concerned itself with the link between labor and health shocks. Adda et al. (2009); Guvenen et al. (2021); Roos et al. (2021) study the effect of labor shocks on health outcomes, while García-Gómez et al. (2013); Lundborg et al. (2015); Lindeboom et al. (2016); Broten et al. (2022); Dobkin et al. (2018) study the effect of health shocks on labor outcomes. The common thread in this literature is that the labor and health domains are intimately related to each other, which we corroborate in this paper. However, this literature tends to take a limited perspective on labor and health shocks by requiring that they meet an exogeneity condition in order to establish causality. Subsequently, methods like difference-in-differences and propensity score matching are employed to examine, for instance, the difference in labor income for comparable individuals who experienced unexpected hospitalization versus those who did not.

Our analysis does not rely on establishing causal relationships, because our aim is to simply characterize the distribution of risks across different domains. As a result, we are less

restricted in the type of shocks that we can consider. In this sense our paper fits in with the literature on prediction policy problems (Kleinberg et al., 2015), where the objective is to predict someone's future state given their current state. Within this literature, machine learning methods are often used to assess individual risks as they are optimized for predictive power due to their capability of incorporating complex interactions and large numbers of variables. One such example is Mueller and Spinnewijn (2023), who study the predictability of long-term unemployment using Swedish administrative data. Another example is Einav et al. (2018), who estimate individual mortality risks.

Mullainathan and Spiess (2017) provide an extensive overview of machine learning applications in economic prediction problems and policy domains, while the OECD study by Desiere et al. (2019) presents an overview of how statistical profiling is already used by governments to predict long-term unemployment and how these predictions are translated into policy actions. We contribute by estimating risks across the domains of labor and health.

In a Dutch context, the studies by Van Hoenselaar et al. (2023) and De Klerk et al. (2023) look at which groups are vulnerable in a conjunction of different domains, such as income, wealth, housing and the personal sphere. The vulnerable groups identified in these studies overlap with those groups we see over-represented in the upper tails of our estimated risk distributions. However, our ex-ante perspective on risks (as opposed to the ex-post perspective on realizations) not only provides a more comprehensive view of the population at risk in contrast to focusing solely on those who have already experienced a shock, but it also sheds light on potential opportunities for preventive policy measures.

3 Data

We rely on the administrative data infrastructure of Statistics Netherlands, which is the Dutch national statistical office, and compile information on the universe of Dutch individuals in three domains: 1) demographic and socioeconomic characteristics, 2) employment, income

and wealth, and 3) healthcare treatments. Statistics Netherlands compiles data from various sources such as municipality population registries, tax returns and health insurance claims, and grants access to the anonymized data for the purpose of scientific research. The analysis is conducted on their servers through a remote access application and the export of results is subject to mandatory compliance inspections to ensure that they meet strict privacy standards. The different data modules can be linked thanks to identifier variables that uniquely pinpoint individuals and households. Table 1 provides further details on the type of information that is available to us in each of the three domains.

The data altogether spans roughly five hundred variables that are available at a yearly frequency. We construct a dynamic panel data set where the unit of observation is a person-year combination. Demographic and socioeconomic characteristics are recorded at the start of the year, employment information is gathered for the highest earning job in that year, and healthcare treatments and costs are added up by broad category throughout the year. We make use of the time period 2013-2018 because the variables underlying the shock definitions as outlined in section 4 are available and consistently measured during this time period. Variables indicating sums of money are deflated to the price level of 2015. Observations with missing values are not omitted from the analysis, but instead missing values are treated as separate informative values by the machine learning models that we train below.[2] It is paramount to not only focus on complete observations, because high-risk individuals are likely to be absent from some of the data modules we use. For example, demographic characteristics have near-universal coverage but the health data is known to be less complete.

In each year, roughly seventeen million individuals are present in the data set, but we shrink the sample in various ways. First, we exclude individuals aged below 25 years or above 60 years. Since our aim is to study the interaction of labor and health risks, we focus on the share of the population that is most likely to be active in the labor force. It is probable that labor and health risks are particularly related as people approach retirement due to

[2]In fact, we find that observations with missing values often show up in the upper tails of our estimated risk distributions.

Table 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.

the possibility of early retirement. We feel that the nature of this interaction is distinctly different from the general inter-dependencies between labor and health risks that we aim to characterize. Second, we exclude individuals with an unknown or uncommon household composition such as student housing or care homes. Third, we require that the values for all shock variables can be calculated for a given person in a given year. This ensures that, for each observation, we obtain risk estimates for all shock definitions so that we can accurately

characterize risk concurrence at the individual level. It requires no missing values and no negative monetary values for the variables underlying the shock definitions. This could introduce a selection effect if individuals with missing values face different levels of risk, but our aim is to characterize the risk distribution of the general population that is well-covered by the national statistical office.

These three sample selection criteria significantly reduce the number of available observations by 55.6%, 1.0% and 16.7%, respectively, but this still amounts to over 25 million observations in total. However, due to computational limitations of the servers at Statistics Netherlands, we randomly select one set of two million observations which we use to train machine learning models and randomly select another set of two million distinct observations which we use to evaluate their prediction performance.

4 Shock Definitions

We study adverse shocks in the domains of labor and health, encompassing a total of twelve distinct shock definitions. To get the main results across, the majority of our analysis is centered around two of these shocks, one for each domain. An overview of all the shock definitions and their respective prevalence within the sample is provided in table 2. The prevalence is defined as the number of shock realizations as a percentage of the number of observations that are eligible to receive the shock. In defining the shocks and corresponding thresholds, we aim for a prevalence approximately between 2.5% and 10%. This approach guarantees that shocks have significant impact and do not occur routinely at the individual level. It makes them relevant for policy makers to act on, but not so scarce that prediction is difficult. Almost all shocks are defined using a precondition, and only those observations that meet the precondition are considered in the analysis. We will explain this in more detail after having introduced the shocks. More detailed information about the variables used in the shock definitions can be found in appendix A.

Table 2: Shock definitions. Prevalence of the shock indicates the number of shock realizations as a percentage of the number of observations that are eligible to receive the shock. The fraction of eligible observations in the sample indicates the share of observations that meet the precondition implied by the shock definition.

Shock	Definition	Prevalence	Eligible
main shocks			
social_benefits	Social benefits become primary source of income.	2.3%	87.7%
health_expenditures	Increase of healthcare expenditures of at least 5,000 euros compared to the year before.	3.6%	100%
alternative labor shocks			
relative_drop_income	Income drop of at least 25% and income of at least 5,000 euros in the year before.	8.9%	81.1%
absolute_drop_income	Income drop of at least 10,000 euros.	8.5%	77.1%
problematic_debt	Individual starts defaulting on health insurance premium payments.	0.5%	97.6%
economic_dependence	Income from labor or entrepreneurship drops below the net social assistance allowance for a single person.	3.9%	72.4%
alternative health shocks			
physical_health_expenditures	At least 5,000 euros of specialized medical care expenditures and at most 1,000 euros in the year before.	2.2%	77.2%
physical_health_treatment	Four or more Diagnosis Treatment Combinations (DBCs) and at most one in the year before.	3.3%	80.3%
physical_health_ic	At least one day on intensive care and none in the year before.	0.3%	99.8%
mental_health_expenditures	At least 2,000 euros of mental health care expenditures and none in the year before.	1.3%	93.9%
mental_health_treatment	At least one mental health treatment process and none in the year before.	1.5%	96.2%
mental_health_medication	Start taking antidepressants, antipsychotics or sedatives.	2.3%	90.7%

There is a wide range of possibilities to consider when constructing the shock definitions. A shock could be measured as the occurrence of a particular event, such as hospitalization or becoming a recipient of social benefits. Alternatively, a shock can be measured indirectly, for example, through an increase in health care expenditures or a drop in income. The advantage of the first, more direct, approach is that it provides a clear definition that does not require a somewhat arbitrary choice of threshold. However, it may prove more challenging to relax the shock definition in cases of low prevalence in the data. For the indirect shock definitions, there is also the consideration of whether to analyze the relative or absolute change of an outcome. For instance, one might consider an income drop of 25% versus a drop of 10,000 euros. The impact of the absolute income drop is much larger for people with low income, whereas the relative change is more significant for individuals with high income. To cover the various possibilities, we have defined shocks in several different ways.

Another consideration is whether to define a shock as a completely new adverse situation or as the worsening of an already existing situation. For example, being a new recipient of social benefits versus continuing to receive them for another year. We focus on new adverse situations to ensure that our shocks define significant turns of events. Furthermore, we limit ourselves to shocks that primarily concern the individual. That is, we do not consider shocks on a household level, such as the job loss of one's partner. In terms of the time dimension of the shocks, we focus on shocks taking place one year ahead, rather than considering shocks that may occur several years in the future.[3]

With the exception of the shock *health_expenditures*, all other shocks have some sort of precondition related to the previous year. For instance, individuals who were already receiving social benefits in year $t-1$ are inherently ineligible to experience the shock *social_benefits* in year t. Similarly, individuals with an income below 10,000 euros in year $t-1$ cannot receive the shock *absolute_drop_income* in year t by construction. Consequently, our prediction model assigns a near-zero risk estimate to the observations that do not meet these

[3]We briefly return to this point in appendix D.4.

preconditions. The inclusion of non-eligible observations would skew our results. Therefore, observations that do not meet the precondition of a shock are excluded from the analysis of that specific shock. This choice comes with a drawback, namely a reduction in sample size. The right column in table 2 indicates the percentage of the total sample that is eligible to receive a shock, i.e., the share of observations that meet the precondition. In the worst case we lose about 27.6% of the observations. Taking an average over the shocks, we lose about 12% of the observations.[4]

For the sake of simplicity, most of the analysis concentrates on two main shocks. These shock definitions offer a high-level perspective on individuals experiencing adverse situations in the domains of labor and health. The alternative shock definitions consider more specific and detailed adverse situations. In the labor domain, the main shock is denoted by *social_benefits*. This shock is defined as the event in which social benefits become an individual's primary source of income in year t, provided this was not the case in year $t-1$. The social benefits considered include unemployment, social assistance, occupational disability, sickness, and other social benefits.[5] This shock has a prevalence of 2.3% in the data set. The main shock in the health domain is denoted by *health_expenditures*. It is defined as an increase of 5,000 euros in health care expenditures from year $t-1$ to year t. This includes expenses for both physical and mental health care, but excludes birth care and general practitioner expenditures. The prevalence of this shock in the data set is 3.6%. It is important to recognize that one's healthcare expenses may not accurately reflect their state of health. People who are in poor health might postpone seeing a doctor for an extended period, and successful treatments can lead to a significant improvement in one's health. Nevertheless, considering the available data, we view it as a satisfactory proxy.

Note that we categorize shocks as adverse events. Yet, various scenarios exist in which a

[4]In section 7 we analyze the concurrence of risks. For the parts of the analysis where two shocks are considered at the same time, each observation has to satisfy the precondition of both shocks. At worst, this decreases the sample size by about 40%, but on average by about 20%.

[5]The components of occupational disability and sickness are inherently also linked to the health domain. Nevertheless, we consider *social_benefits* a labor shock, as it represents a scenario in which an individual's primary source of income no longer stems from their employment or business.

shock represents a deliberate choice rather than an unexpected event. Consider, for instance, the voluntary decision to reduce working hours in favor of allocating more time to family or other personal activities. Unfortunately, the data does not allow us to differentiate between setbacks and proactive, positive decisions.

5 Risk Predictability

In section 4 we defined the conditions for an individual to be said to have experienced a shock in a certain year. This gives us, for each individual i in each year t, a variable $s_{i,t}$ that expresses the realization of the shock. This variable $s_{i,t}$ takes either the value 1, for a shock realization, or the value 0, for no shock realization. What this binary realization fails to accurately capture is the underlying probability that individuals had of facing a shock realization. It is this probability that we are actually interested in, since it tells us how much at risk different individuals are. To this end, we employ machine learning methods. The objective of this approach is to identify for each individual i in each year t a variable $p_{i,t}$ that expresses the probability that individual i will have a shock realization in year t. We can achieve this by asking a machine learning model to predict shock realizations for individuals based on their data from previous years, assigning a score between 0 and 1. We may interpret these scores as a representation of the underlying probabilities $p_{i,t}$ of an individual i having a shock realization in year t, provided that our machine learning predictions are accurate. In this section we will describe the specific machine learning method we use and show that our prediction models produce unbiased estimates of the true underlying shock probabilities.

5.1 Machine Learning Method

We use the *R* package *LightGBM*, which implements gradient boosting machine learning methods. This means that our prediction models are an ensemble of simpler models (in our case, decision trees). We train gradient boosted trees models using a set of data points that

we call our train set, and then subsequently have it make predictions on a different set of data points that we call our test set. For the full set of parameters that were used when training the models, see appendix B.

Each data point in our train and test sets represents the observation of a single individual in a single year. The train set consists of 2,016,862 such observations based on 1,074,640 individuals, and the test set consists of 2,008,969 such observations based on 1,074,640 different individuals.[6] For each individual, we have both time-invariant and time-dependent variables. Of the time-dependent variables, we include three lags of the variable in each observation. Together with the time-invariant variables, this results in 1,283 variables for each observation. These variables include categorical and missing values, both of which are conveniently handled by the *LightGBM* package. To avoid data leakage, an observation includes no time-dependent information relating to the year for which the prediction is made.

For each shock definition, we train one single model to make predictions for all observations in the test set. The train and test sets contain observations from across all years that we have data for. Note that year fixed effects can be picked up by the model through the year variable.[7] Also note that our method of constructing the train and test sets guarantees that the observations of any individual are either all in the train set or all in the test set.

5.2 Prediction Evaluation

Given a particular shock definition, we have the shock realization $s_{i,t}$ and the risk estimate $p_{i,t}$ for each observation in the test set. To assess prediction quality, we would ideally compare our risk estimates to the true underlying risks, but the latter are unobserved. We only observe shock realizations, but these are not directly comparable to our risk estimates. For example, it is unclear whether a risk estimate $p_{i,t} = 0.6$ is accurate, because even though a shock realization of 1 is more likely, one would expect a realization of 0 in about 40% of

[6] The slight size difference in the train and test set arises because we first randomly select an equal number of individuals and then apply the sample selection criteria as described in section 3.

[7] Because we do not observe substantial differences in shock prevalence throughout the sample period of 2013-2018, no significant year fixed effects are expected.

cases. This means that aggregation over individual risk estimates and shock realizations is required. Therefore, we compare the realized prevalence to the mean risk estimate for groups of observations.[8]

We can assess the prediction quality by regressing the shock realizations on our risk estimates. If the machine learning model produces unbiased risk estimates, the resulting regression should trace the 45° line. Figure 1 shows that this is indeed the case for the two main shocks *social_benefits* and *health_expenditures* with the intercept being very close to 0 and the coefficient being very close to 1. This implies that the random realization of the underlying risks averages out on an aggregated level and that there are no significant shifts in the true underlying risks between the moment of prediction and realization.

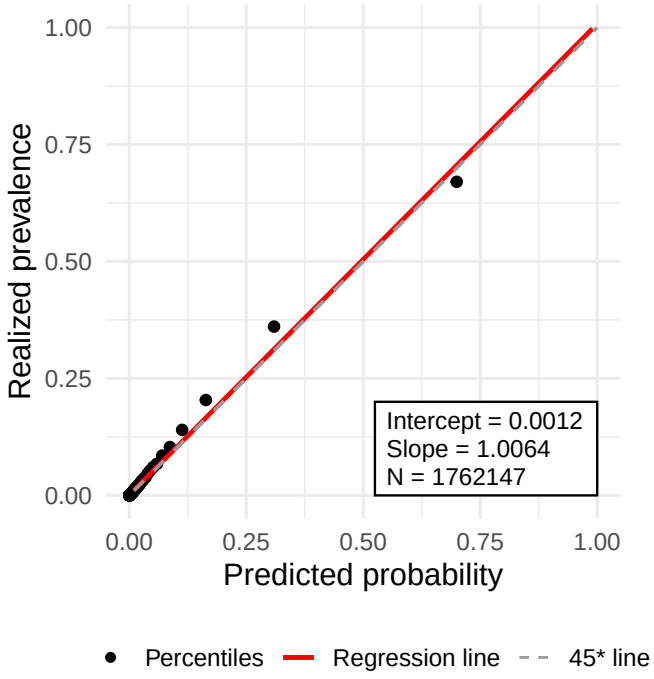
The regression line represents an aggregation at the level of the entire data set, but we can evaluate the predictions throughout the entire estimated risk distribution by dividing our test set into percentile bins of relative risk estimates. We know the mean risk estimate and the realized prevalence among the approximately twenty thousand observations in each bin. An unbiased machine learning model should produce mean risk estimates that are close to the realized prevalence for all bins. The scatter points in figure 1 show that this is indeed the case.

The R-squared of the regression in figure 1 indicates how much of the variation in the actual realizations is predicted by the ex-ante risk estimates. A higher R-squared implies that the realization of an adverse event carries a larger degree of predictability. Table 3 reports the R-squared for all twelve shock definitions. It is evident that a hierarchy of predictability emerges. Labor shocks are the most predictable, followed by mental health shocks and physical health shocks. For reference, Mueller and Spinnewijn (2023) obtain an R-squared of 0.136 when predicting job finding rates with similar methodology.

[8]The alternative is to collapse our continuous individual risk estimates to binary realization predictions by choosing a classification threshold. The prediction evaluation of this classification exercise is presented in appendix C.

Figure 1: Regressing ex-post realizations on ex-ante risk estimates, including a binned scatter plot of the shock prevalence (on y-axis) and the mean risk estimate (on x-axis) for each percentile of the risk distribution.

(a) *social_benefits*



(b) *health_expenditures*

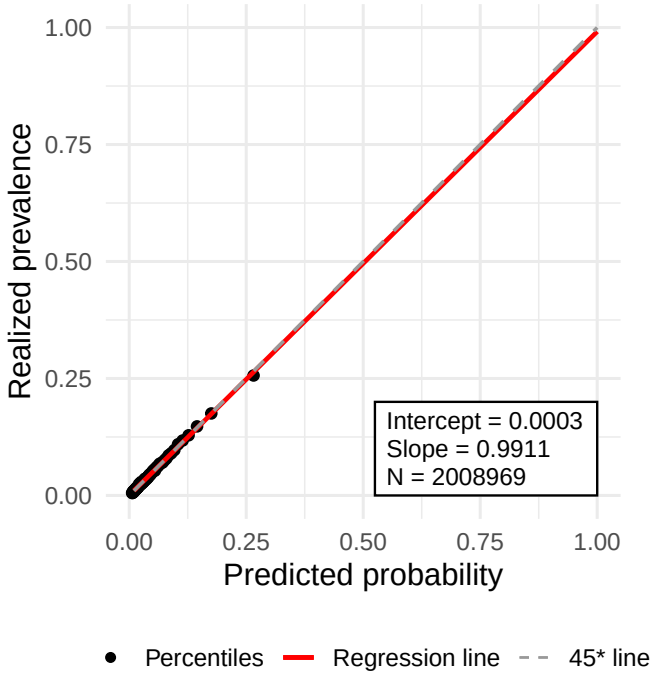


Table 3: R-squared of regressing ex-post realizations on ex-ante risk estimates.

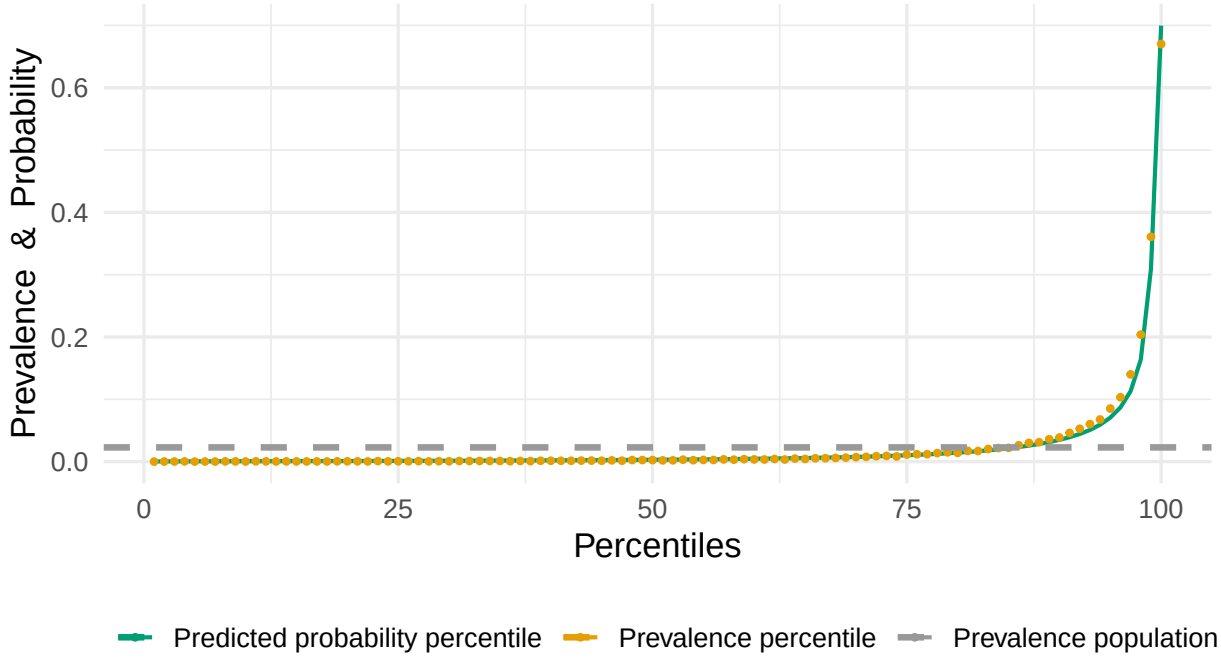
Shock	R-squared
social_benefits	0.290
health_expenditures	0.043
relative_drop_income	0.254
absolute_drop_income	0.262
problematic_debt	0.030
economic_dependence	0.278
physical_health_expenditures	0.018
physical_health_treatment	0.044
physical_health_ic	0.003
mental_health_expenditures	0.035
mental_health_treatment	0.051
mental_health_medication	0.043

6 Risk Distributions

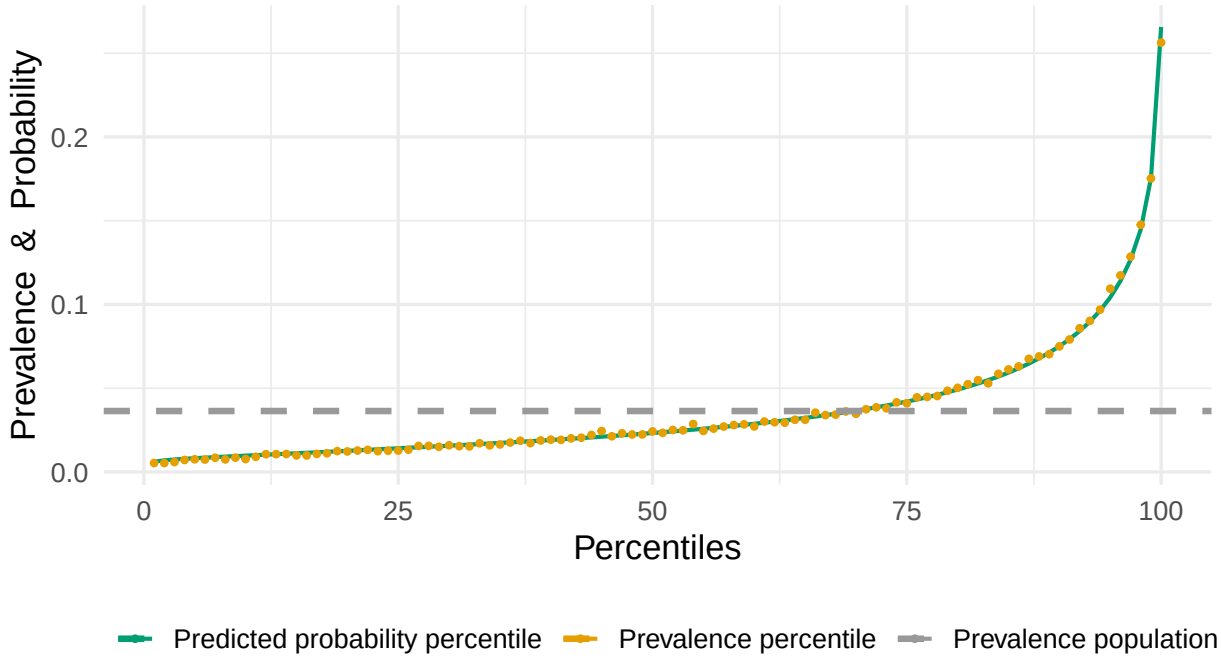
We established in section 5 that we obtain unbiased risk estimates throughout the entire distribution, but we did not yet discuss the shapes of the risk distributions. Because most scatter points are bunched together in figure 1, we present an amended version in figure 2 with the mean risk estimates (green line) and realized prevalences (orange dots) across the percentiles on the x-axis. We see that our model is able to discern who bears the most risk, as evidenced by the sharp increase in both the risk estimate and realized prevalence towards the upper tail. Our model is thus able to pick out a large group of people that faces substantially heightened risks of setbacks on which it accurately matches the realized prevalences. It is evident that there is a lot of risk inequality throughout the population, especially for the labor shock. Compared to the population average (grey stripes), the highest risk percentile faces a risk that is up to thirty times (in the case of main shock *social_benefits*) and ten times larger (in the case of main shock *health_expenditures*). Furthermore, over two thirds of the population faces below average levels of risk.

Figure 2: Risk distribution. The y-axis depicts the mean risk estimate (green line) and realized prevalence (orange dots) for each percentile of the risk distribution.

(a) *social_benefits*



(b) *health_expenditures*



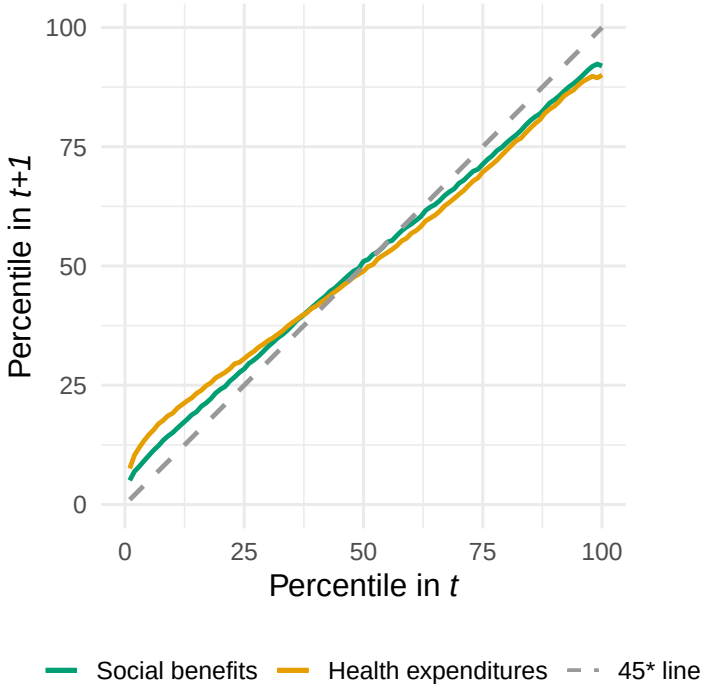
It is unclear from the static distribution of risks in figure 2 whether the same people end up in the same regions of the risk distributions over time. Figure 3 shows that one's rank in the risk distribution is indeed very persistent. For example, the people that are in the highest percentile of the *social_benefits* risk distribution in year t are on average in the 92^{nd} percentile in year $t+1$, provided that they did not receive a shock realization in that year. The persistence is not only present for the two main shocks in figure 3(a), but also for the group of combined labor shocks, mental health shocks and physical health shocks in figure 3(b). Generally, the persistence is strongest for labor shocks followed by mental health shocks and physical health shocks, but mental health shocks are most persistent in the upper tail of the risk distribution.[9] The widespread persistence suggests that our risk estimates represent individual fixed effects or one's ex-ante latent risk type.

We have experimented along various dimensions to gain additional insights about the predictability and distribution of risks. These analyses can be found in appendix D. In appendix D.1 we take a look at variable importance, in appendix D.2 we discuss the predictions made by a machine learning model trained on a data set with only a few rudimentary variables, in appendix D.3 we discuss the predictions that arise when we artificially make the shock prevalence more balanced, and in appendix D.4 we discuss alternative shock definitions.

[9]The dip of labor shock persistence in the upper percentiles of figure 3(b) is due to a selection effect in the *relative_drop_income* and *absolute_drop_income* shocks. Most of the individuals faced income drops that did not meet the shock threshold and in turn were less likely to receive the shock the next year.

Figure 3: Rank persistence of risk estimates. The y-axis depicts the average rank in the risk distribution during the following year given one's rank in the current year.

(a) *social_benefits* and *health_expenditures*



(b) Average of shocks in the domain of labor, mental health and physical health

7 Risk Concurrence

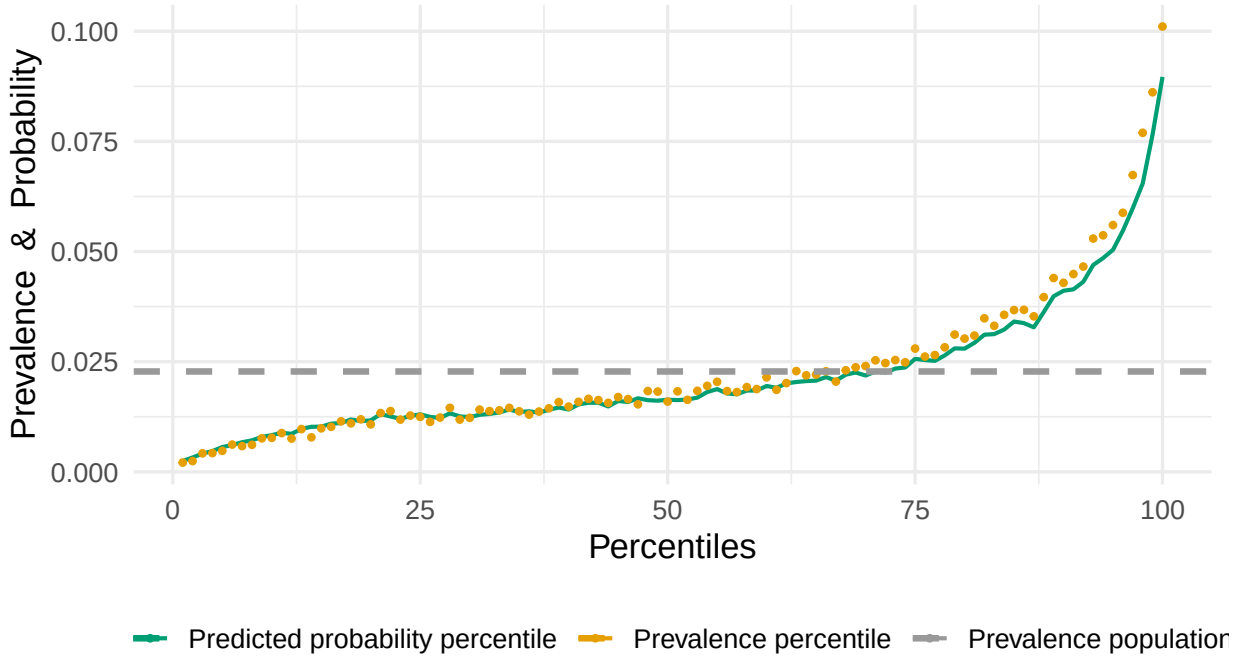
Having characterized the risk distributions of individual shocks, we now move on to study how the risk distributions of different shocks relate to each other. We will show that there is strong concurrence between risks not only within a particular domain, but also across domains. Individuals with a high estimated risk for one particular shock are also more likely to face different shocks. Furthermore, we will highlight that this piling up of risks is substantial and monotone throughout the entire distribution of risks.

7.1 Risk Distributions across Domains

The figures of section 6 can be amended to contrast the risk distributions across domains. For example, figure 4 depicts the joint risk distribution of the two main shocks, where for each percentile of the *health_expenditures* risk distribution we plot the average predicted and average realized risk of the *social_benefits* shock. Evidently, there is a strong positive association between the risks that individuals face. This not only holds for the ex-ante predicted risks (green line), but also for the ex-post realizations (orange dots). Our machine learning model is thus not only able to identify who is most likely to be hit by an adverse event, but also able to identify who is most likely to be confronted with the joint materialization of adverse events.

People with the lowest chance of a sharp increase in health expenditures also face the lowest probability of having to suddenly rely on social benefits, and vice versa. The inequality in risk exposure is substantial: those in the upper tail of the *health_expenditures* risk distribution are up to four times more likely to be confronted with the *social_benefits* shock compared to the population average and ten times more likely compared to the lower tail of the *health_expenditures* risk distribution. Another striking fact is that the association between the two risks is monotone. Risks are not only correlated in the upper tail of the distributions, but instead the correlation is present throughout. A one percentile jump

Figure 4: Joint risk distribution of *health_expenditures* (on x-axis) and *social_benefits* (on y-axis).

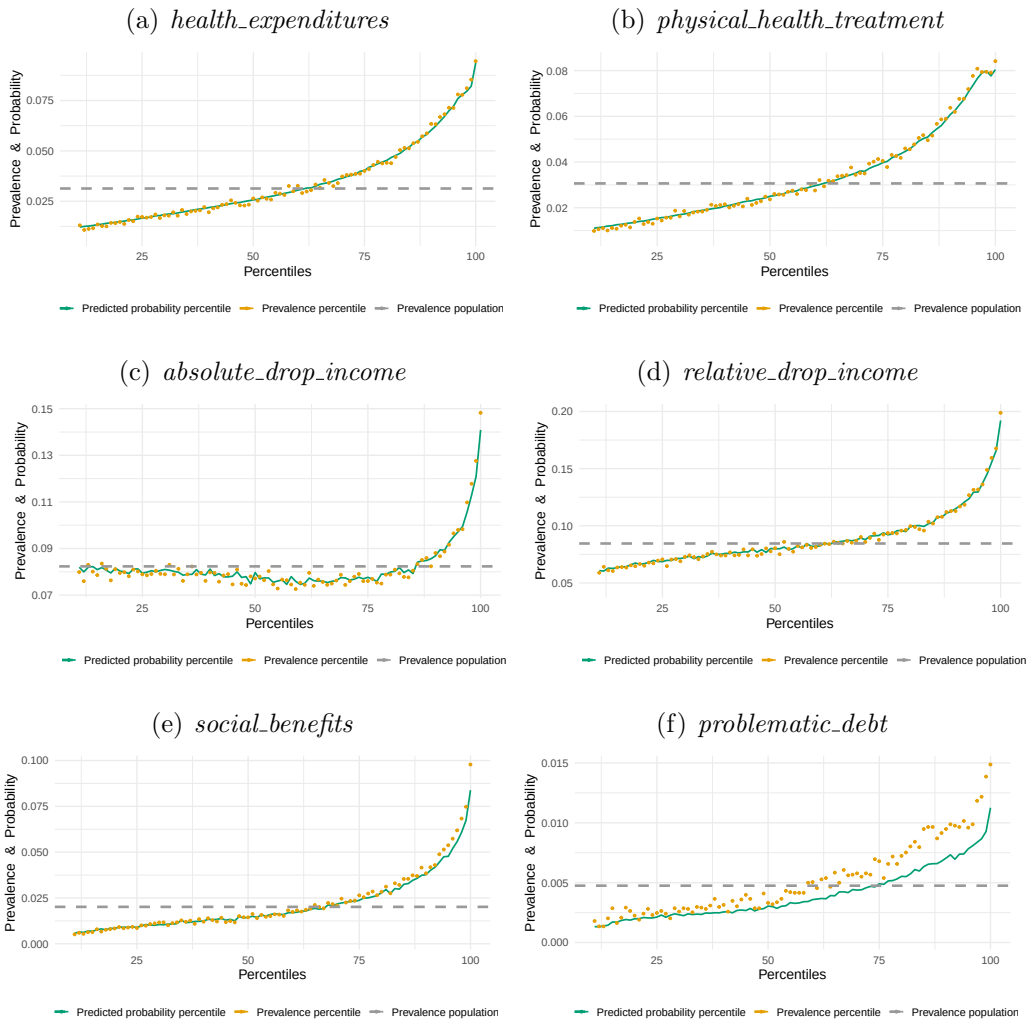


up in the risk distribution of *health_expenditures* implies a jump up in the risk distribution of *social_benefits* as well, regardless of the initial position in the *health_expenditures* risk distribution.

These novel insights become visible thanks to our focus on ex-ante risks. With ex-post realizations one could only arrive at a 2×2 matrix containing the joint counts of 0's (no shock materialization) and 1's (shock materialization) for both shocks, and one could calculate a correlation coefficient to quantify the coincidence of both shocks. However, these realizations are influenced by risk type heterogeneity as well as by bad fortune. The resulting correlation coefficient would not answer to what extent shock coincidence is due to either factor. Our approach disentangles the two, uncovers the full distribution of risks, and allows to quantify the dispersion of risks borne by individuals and the monotonicity between risks across domains.

To highlight that the concurrence of risks not only materializes for our main shocks but that it is a persistent feature of all the shocks that we consider, we show the joint risk

Figure 5: Joint risk distribution of *mental_health_medication* (on x-axis) and various other shocks (on y-axis).



distribution of *mental_health_medication* with various other shocks in figure 5. While a positive association between the risk of consumption onset of anti-depressants, anti-psychotics or sedatives and the risk of a sizable increase in general health expenditures (figure 5(a)) is not surprising since the former is a nested category within the latter, it is interesting to observe a positive association with the risk of a sizable number of physical healthcare treatments (figure 5(b)). Apparently, even within a particular domain such as health, there is cross-over in risk between subdomains such as mental health and physical health. Moving

on to the cross-domain joint risk distributions, we observe a positive association between *mental_health_medication* and the risk of a sizable absolute or relative drop in labor income (figures 5(c) and 5(d))[10] and in terms of having to suddenly rely on social benefits (figure 5(e)). This extends from the labor domain into the wealth domain, where the labor income fragility comes with a heightened likelihood of defaulting on mandatory health insurance premium payments (figure 5(f)), which is a commonly used indicator for the onset of problematic debt in the Netherlands.

Taken together, figure 5 illustrates significant concurrence of mental health, physical health, labor income and problematic debt risk, where individuals in the upper tail of one risk distribution are highly likely to be present in the upper tail of many other risk distributions at the same time, and where individuals in the lower tail of one risk distribution face relatively little risk at all. A similar narrative emerges when inspecting figures E1 and E2, where we show additional joint risk distributions of the two main shocks with shocks from the other domain.

A brief discussion of the interpretation of these results is in order. Our view on risk concurrence does not require that the realization of one shock directly causes the materialization of another shock. In fact, it is probable that single events have multi-faceted consequences that are drawn out over time. We would then register both initial and subsequent shock realizations even though they all share the same root cause. If this is how most setbacks permeate, then our results show that setbacks have widespread and far-reaching implications if one takes into account cross-overs between domains. We stated above that our predicted probabilities are estimates of the latent type heterogeneity that is present for each shock. The strong degree of risk concurrence suggests that these are themselves estimates of another latent variable: the susceptibility to setbacks in general.

[10]Individuals with a high risk of *mental_health_medication* have lower incomes and are therefore less likely to face an income decline exceeding 10,000 euros. If we only applied the eligibility condition for *mental_health_medication* and not for *relative_drop_income*, the figure would show a clear negative association between the risk estimates.

7.2 Correlation of Risk Estimates across Domains

In the previous subsection we depicted a curated set of joint risk distribution plots, because it is infeasible to display them for all possible combinations of the twelve shocks that we consider. To underscore the ubiquity of risk concurrence, we calculate Spearman rank correlation coefficients between the risk estimates of all shock pairs and present them in table 4. While the commonly used Pearson correlation coefficient assesses the linearity between two variables, the Spearman rank correlation coefficient instead assesses the monotonicity between two variables. The figures in section 7.1 show that the relationship between two risk distributions tends to be monotone, but not necessarily linear. This exercise collapses each pair of risk distributions to one number that measures their degree of association. It succinctly condenses two important insights.

First, all values have a positive sign, which indicates that the positive association found in the plots of section 7.1 is also present in all other shock combinations. Second, the association between risk estimates is fairly monotone for most shock pairs. To get a feeling for what degree of monotonicity each value in table 4 represents, we map a subset of numbers to the earlier presented joint risk distribution plots. Figure 4 amounts to a Spearman rank correlation coefficient of 0.31, figure 5(b) to a coefficient of 0.70, and figure 5(c) to a coefficient of 0.01. The second value signals that the risk concurrence between *mental_health_medication* and *physical_health_treatment* is spread evenly throughout the joint risk distribution, while the latter value shows that the risk concurrence between *mental_health_medication* and *absolute_drop_income* is mostly concentrated in the upper tail. The top-left and bottom-right quadrants of table 4 imply that the risk concurrence within a domain is particularly monotone.[11] Furthermore, the bottom-left quadrant highlights that the risk concurrence between the labor and mental health domain is more monotone than that between the labor and physical health domain.

[11] This stems also from the fact that the shock definitions in each domain measure similar concepts.

Table 4: Correlation matrix. This table displays the Spearman rank correlations between the risk estimates of all different shocks. Due to the large sample size, all correlation coefficients are different from zero with high statistical significance.

	social_benefits	*relative_drop_income*	*absolute_drop_income*	*problematic_debt*	*economic_dependence*	*health_expenditures*	*physical_health_expenditures*	*physical_health_treatment*	*physical_health_ic*	*mental_health_expenditures*	*mental_health_treatment*
relative_drop_income	0.60										
absolute_drop_income	0.36	0.76									
problematic_debt	0.44	0.32	0.05								
economic_dependence	0.66	0.83	0.46	0.42							
health_expenditures	0.31	0.17	0.07	0.21	0.22						
physical_health_expenditures	0.15	0.08	0.00	0.12	0.15	0.89					
physical_health_treatment	0.15	0.09	0.04	0.06	0.12	0.86	0.88				
physical_health_ic	0.11	0.07	0.09	0.14	0.06	0.63	0.50	0.48			
mental_health_expenditures	0.51	0.25	0.11	0.39	0.30	0.55	0.32	0.39	0.14		
mental_health_treatment	0.53	0.28	0.15	0.41	0.32	0.62	0.37	0.44	0.23	0.93	
mental_health_medication	0.40	0.22	0.01	0.36	0.31	0.78	0.64	0.70	0.44	0.69	0.75

7.3 Conditional Risk Estimates across Domains

One might wonder whether individuals who have recently experienced one adverse event are substantially more likely to be faced with another shock compared to the unconditional probability of the entire population. Table 5 depicts the factors by which the conditional probabilities are larger than the unconditional probabilities based on risk estimates, while appendix table E1 depicts the values based on actual shock realizations. It is clear that the multiplicative factors in both tables are similar, indicating again that our machine learning are able to identify which individuals are truly at risk by producing unbiased risk estimates. Both tables contain no values below 1, underscoring again how pervasive the concurrence of risks is.

Furthermore, the asymmetry between the multiplicative factors in the top-right quadrant and those in the bottom-left quadrant shows that labor risks are particularly elevated after the materialization of health shocks compared to the elevated health risks after the materialization of labor shocks. For example, the predicted probability of the *social_benefits* shock conditional on the *health_expenditures* shock is 2.6 times higher than the unconditional prevalence, while vice versa it is only 1.6 times higher. This suggests that the concurrence of risk flows predominantly from health risks to labor risks. Regardless of this asymmetry, the time ordering in the tables shows that individuals are likely to experience chains of adverse events, where one shock is followed over time by the next, both within and across domains.

Table 5: This table displays the multiplication factors of the average predicted probability in year t for individuals that experienced a different shock in year $t-1$, relative to the unconditional average predicted probability of experiencing that shock in year t.

		conditional on shock in t-1:											
shock in t:	unconditional probability (%)	*social_benefits*	*relative_drop_income*	*absolute_drop_income*	*problematic_debt*	*economic_dependence*	*health_expenditures*	*physical_health_expenditures*	*physical_health_treatment*	*physical_health_ic*	*mental_health_expenditures*	*mental_health_treatment*	*mental_health_medication*
social_benefits	2.3		4.0	3.2	3.4	4.1	2.6	1.9	1.7	3.1	3.6	3.6	2.9
relative_drop_income	8.9	4.0		2.3	2.2	3.1	1.8	1.5	1.4	2.0	2.1	2.1	1.8
absolute_drop_income	8.5	3.0	2.2		1.6	1.7	1.6	1.4	1.3	1.9	1.9	1.9	1.6
problematic_debt	0.4	3.2	2.1	1.6		2.5	1.4	1.2	1.1	1.9	2.0	2.1	1.8
economic_dependence	3.8	7.8	4.3	2.8	2.9		2.0	1.6	1.5	2.4	2.6	2.6	2.2
health_expenditures	3.6	1.6	1.1	1.0	1.3	1.2		1.4	1.9	2.9	1.8	1.9	2.0
physical_health_expenditures	2.2	1.3	1.1	1.0	1.1	1.1	1.6		2.0	3.3	1.3	1.4	1.5
physical_health_treatment	3.3	1.3	1.1	1.0	1.1	1.1	2.7	3.0		3.7	1.7	1.7	1.9
physical_health_ic	0.2	1.7	1.2	1.1	1.5	1.3	5.0	3.3	3.1		2.4	2.4	2.2
mental_health_expenditures	1.3	2.1	1.4	1.1	1.9	1.5	1.8	1.5	1.8	1.9		11.6	3.5
mental_health_treatment	1.5	2.2	1.4	1.2	2.0	1.6	2.5	1.6	1.8	2.4	9.3		4.0
mental_health_medication	2.3	2.0	1.3	1.1	1.8	1.5	2.9	1.9	2.1	3.0	5.4	5.1	

8 Risk Groups

Having constructed risk estimates on an individual level, we study the characteristics of the people in the upper tail of the risk distributions by comparing them to those who are not. This will indicate whether specific person characteristics are correlated with high-risk status and will pinpoint to specific groups that are vulnerable.[12] More specifically, we focus on the features in the previous year of those individuals with the top 5% of risk estimates. This percentage roughly corresponds to the mean shock prevalence observed in the data set (see table 2). Figures 6 and 7 display the distribution of characteristics related to employment, education, wealth, and income, while figures 8 and 9 concern more personal characteristics such as gender, age and marital status. We will briefly comment on the findings of each characteristic.

Contract type, employment type and socioeconomic category. The share of individuals with a fixed-term employment contract (as opposed to a permanent one) is about 2.5 times higher for the group with a high risk of experiencing labor shock *social_benefits* compared to the remaining 95%. For health shock *health_expenditures* the difference between the two groups is marginal. Additionally, for both the labor and health shock, the share of individuals with a part-time contract is about 1.3 times larger. Lastly, the share of self-employed individuals is substantially smaller for both the labor and health risk distribution.[13]

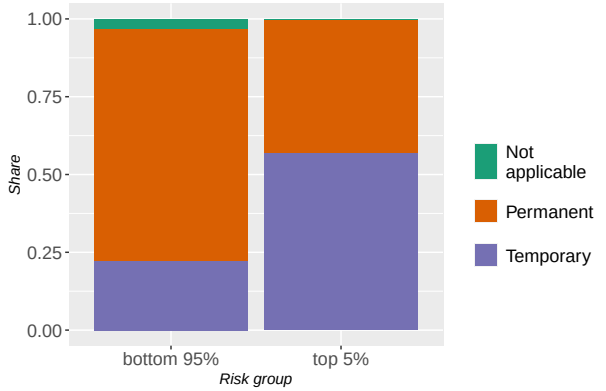
Educational attainment. For both the labor and health shock, the share of people with a low educational attainment (primary school, lower levels of practical education) is about twice as big in the high-risk tail. The share of individuals with a high educational attainment (university, higher professional education) is about half the size in the upper tail of the labor and health risk distribution.

[12] Similar results arise with an analysis based on realizations rather than predictions.

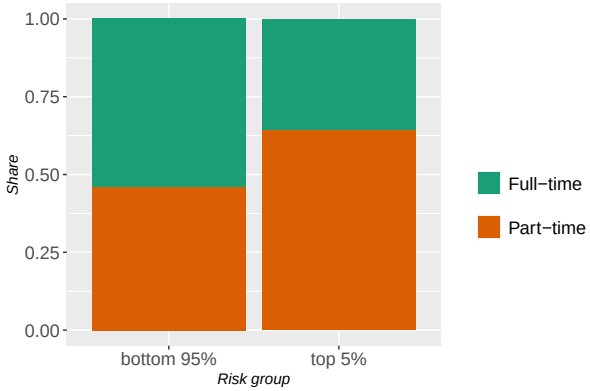
[13] The category 'Other' in these graphs includes people on social benefits and early pensioners. By construction of the shock definition *social_benefits*, individuals who are already on social benefits in year $t-1$ cannot get a shock realization in year t. Since these individuals are excluded for this shock, the category 'Other' is quite small for *social_benefits*. As this precondition is not assumed for the shock *health_expenditures*, the category 'Other' is over-represented in the tail of that shock.

Figure 6: Risk groups for *social_benefits* shock by employment, income and wealth characteristics.

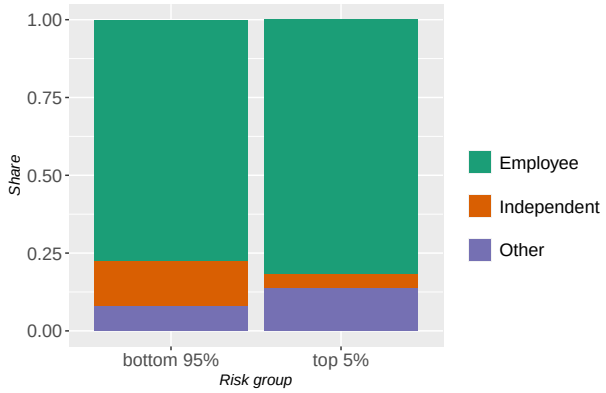
(a) Contract type



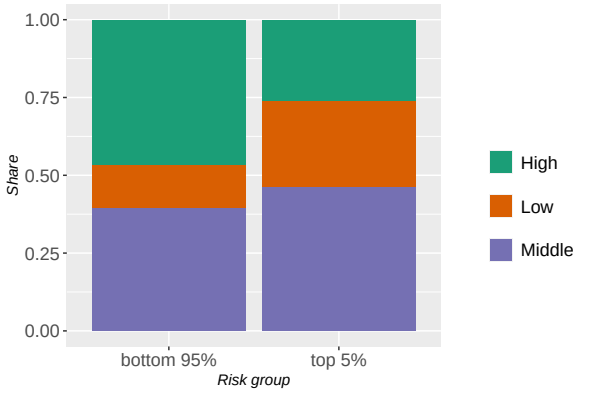
(b) Employment type



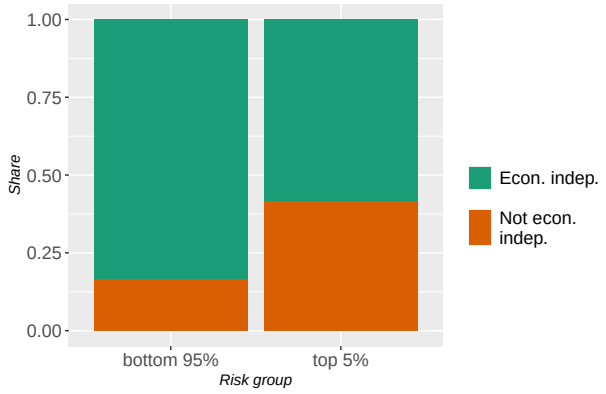
(c) Socioeconomic category



(d) Educational attainment



(e) Economic independence



(f) Wealth percentile

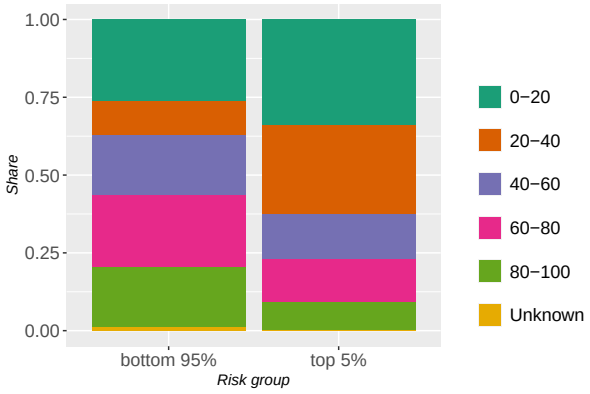
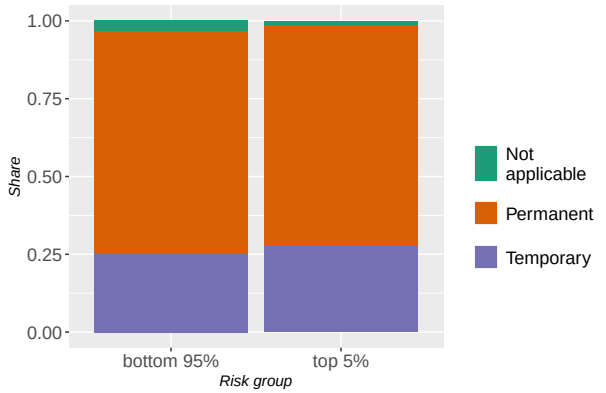
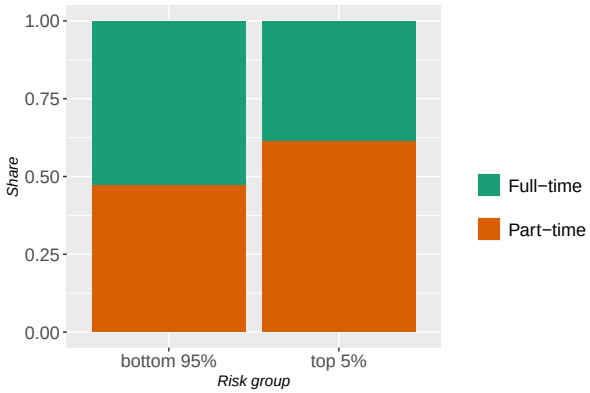


Figure 7: Risk groups for shock *health_expenditures* by employment, income and wealth characteristics.

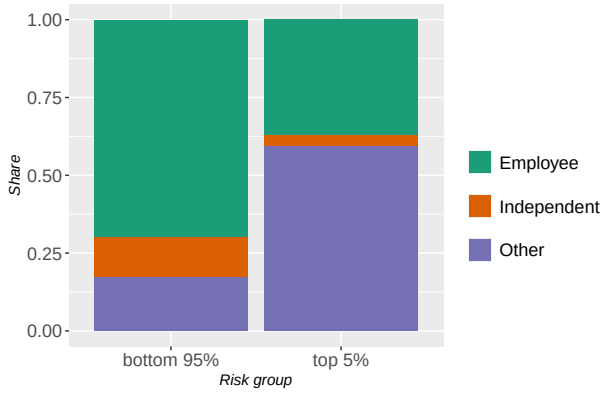
(a) Contract type



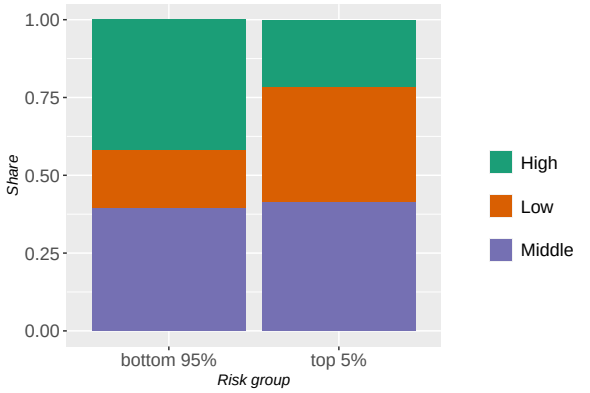
(b) Employment type



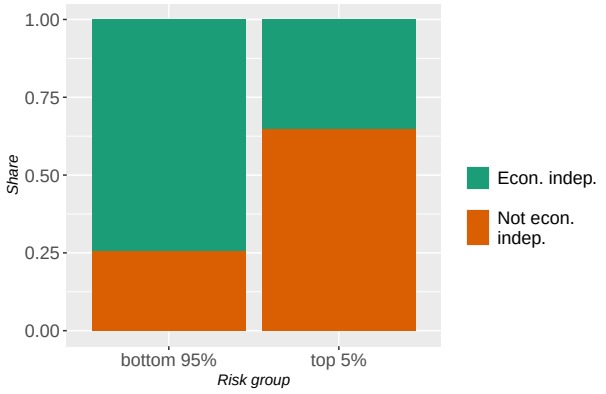
(c) Socioeconomic category



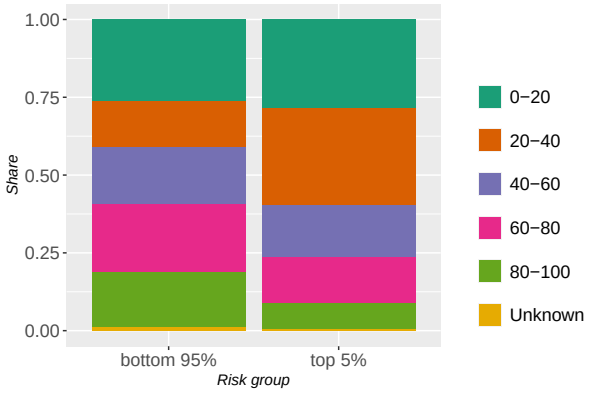
(d) Educational attainment



(e) Economic independence



(f) Wealth percentile



Economic independence and wealth. Economic independence in this context is defined as having a personal income that is higher than 70% of the minimum net wage, which corresponds to the social assistance amount for a single person. The share of people who are not economically independent is about 2.5 times higher in the tail of *health_expenditures*. Because the individuals who are already on social benefits are excluded from the shock *social_benefits*, the level in both the bottom 95% and top 5% is much smaller for *social_benefits*. Additionally, there is an over-representation of the 40% of individuals with the smallest wealth.[14] The shift is especially substantial for the second quintile and less for the bottom quintile of the wealth distribution. For *social_benefits*, this could be explained by the construction of the labor shock, which excludes individuals who are already on social benefits and are presumably over-represented in the lowest wealth quintile.

Gender, country of origin and birth cohort. Approximately 60% of the people in the upper tail of the health risk distribution are women. For the labor shock, the imbalance is much smaller. Furthermore, we see an over-representation of individuals born outside the Netherlands in the upper tail of the labor risk distribution. For the health shock, there are slightly fewer people born abroad that appear in the high-risk tail. There is a small over-representation of young people (birth cohort 1983-1992) in the upper risk tail of *social_benefits*. As expected, older people are over-represented in the upper risk tail of *health_expenditures*.

Housing. There are about four times as many individuals that live in a rental home with housing benefit[15] in the upper tail of the risk distributions of both shocks compared to the rest of the distribution. Albeit to a lesser extent, individuals in rental homes without housing benefit also appear more often in both risk tails.

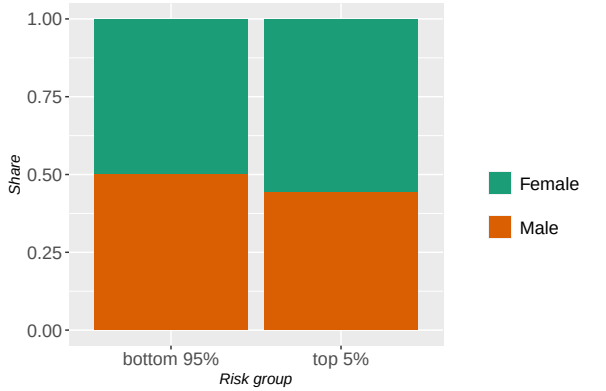
Marital status and household composition. There are roughly twice as many divorced people in the upper risk tails of *social_benefits* and *health_expenditures* compared to the rest.

[14] Wealth here is defined as the household's assets (bank account balances, stocks and bonds, real estate, privately owned firms, and miscellaneous) minus its liabilities (mortgage, student, and other debt).

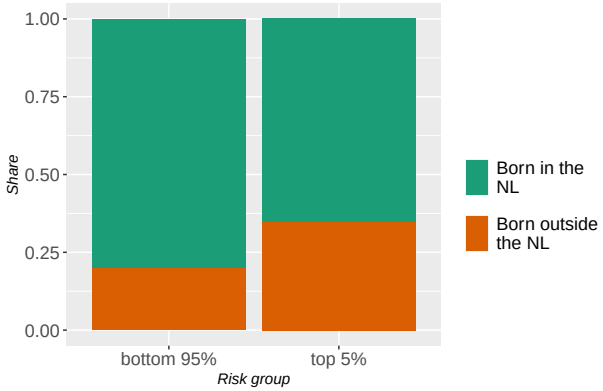
[15] In the Netherlands, housing benefit provides financial assistance to individuals with low incomes who rent their homes.

Figure 8: Risk groups for shock *social_benefits* by a selection of personal and household characteristics.

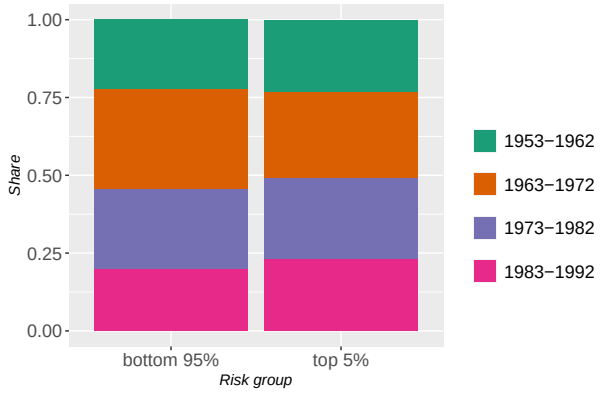
(a) Gender



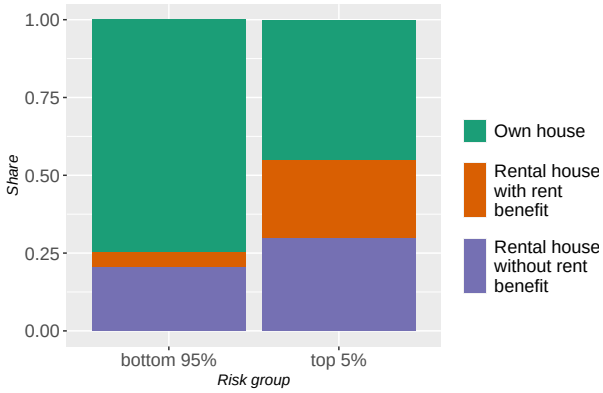
(b) Country of origin



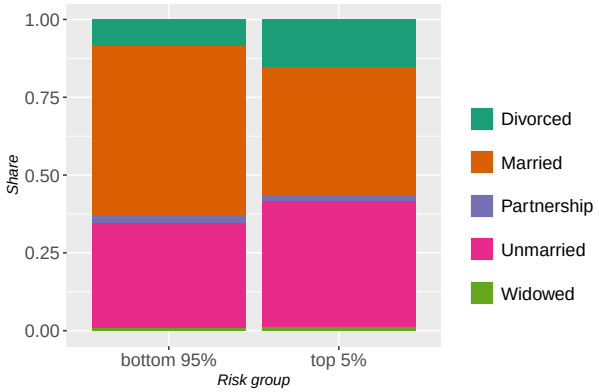
(c) Birth cohort



(d) Housing



(e) Marital status



(f) Household composition

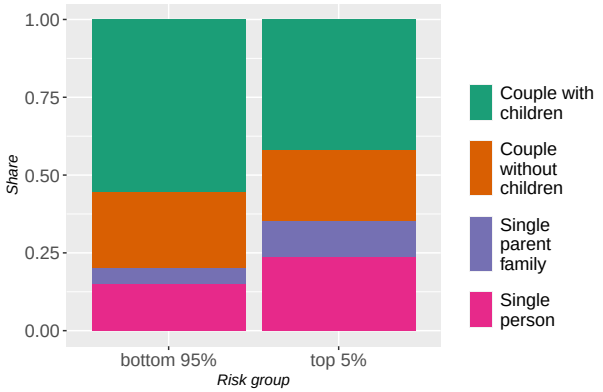
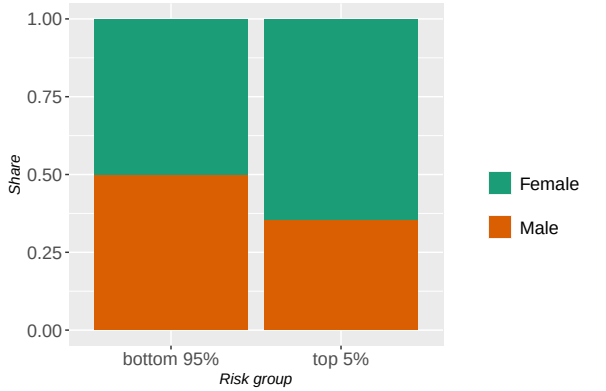
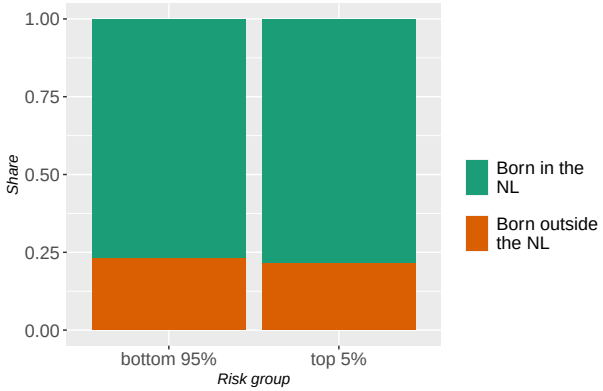


Figure 9: Risk groups for shock *health_expenditures* by a selection of personal and household characteristics.

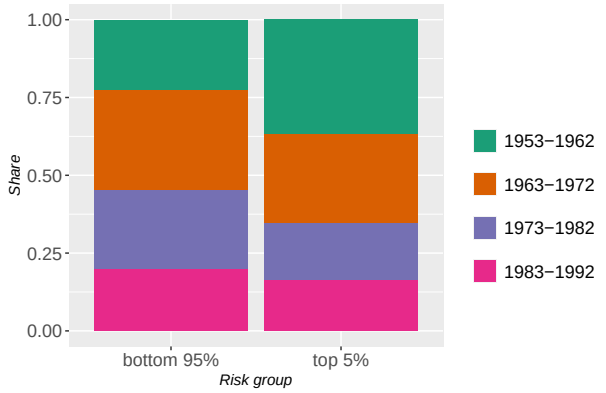
(a) Gender



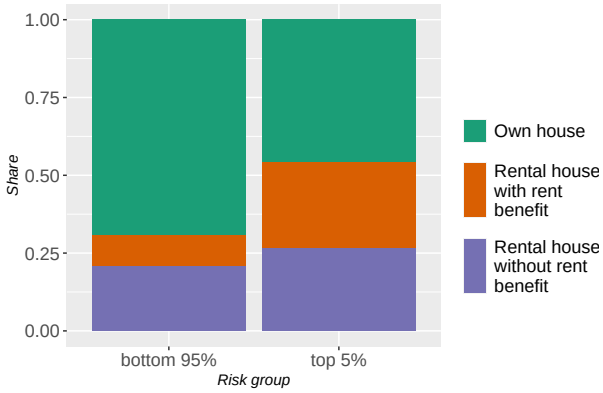
(b) Country of origin



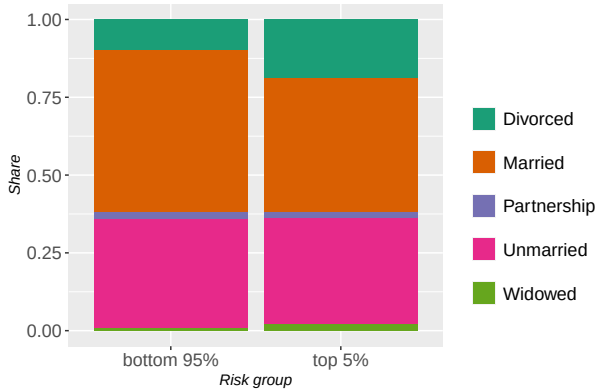
(c) Birth cohort



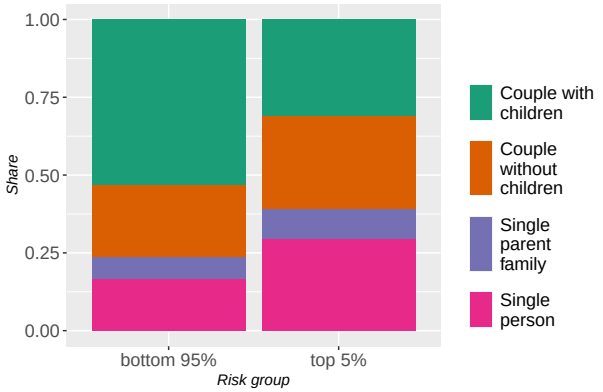
(d) Housing



(e) Marital status



(f) Household composition



Looking at the household composition, we see considerably more singles in the upper risk tail of both *social_benefits* and *health_expenditures*. Additionally, the share of single parent families is approximately twice as large in the upper risk tail of *social_benefits*, and we also see an increase for *health_expenditures*.

A natural extension of this analysis is the exploration of the characteristics shared by individuals who are in the highest risk group for both shocks simultaneously. The results are displayed in figures E3 and E4. The characteristics that stand out in this context exhibit substantial overlap with those identified for singular shocks.

This analysis highlights the unequal distribution of risks within the domains of labor and health across the Dutch population. Specific groups significantly dominate the high-risk end of the distribution. Notable among these groups are individuals with temporary employment contracts, those with lower levels of education, individuals originating from outside the Netherlands, residents of rental properties, and single individuals.[16]

9 Policy Implications

The shape of the risk distributions that we uncover is influenced by existing government policies and economic forces. Policy makers can thus control how risks are spread out across individuals and domains. Risk inequality could be reduced by mitigating the risks at the top of the distributions. Our findings show that it is possible to identify these individuals without having to wait until setbacks occur. This implies that, should effective prevention programs exist, they could receive targeted treatment beforehand. Such proactive policy has broad benefits in the presence of risk concurrence as it could also prevent chains of adverse events. Although predictability of setbacks is not a sufficient condition for targeted prevention policies to be of added value, it is a necessary condition.

Our results help identify sub-populations of manageable size and high risk. This can be

[16]However, the mere over-representation of a particular characteristic in the upper tail of the risk distributions does not imply that having this characteristic automatically signals high-risk status. This is confirmed by the analysis in appendix D.2.

achieved either directly by using the risk estimates from a machine learning model trained on a rich dataset or indirectly by targeting based on person characteristics that correlate with those risk estimates. Tables E2 and E3 consolidate the efficacy of these sources for targeted prevention policies in the domains of labor and health. The machine learning model offers superior predictive power since no combination of person characteristics achieves a higher prevalence for a given size of sub-population. However, it also comes with substantial data requirements and practical challenges (such as privacy concerns), which limit its use as a policy tool. Policy makers may consider targeted prevention policies using socioeconomic indicators or past shock incidences instead. Tables E2 and E3 illustrate that these factors achieve reasonable risk separation, especially when combined. This presents a viable alternative to the fully trained model.

10 Conclusion

In this paper we investigate the predictability and distribution of risks across the labor and health domains. We train machine learning models With extensive data on millions of Dutch individuals and use them to predict individual shock probabilities for twelve different shocks. These risk estimates signify the latent ex-ante risk type of individuals. We verify that our prediction models produce unbiased estimates and document that there is significant risk exposure inequality between individuals. Moreover, we show that risks are correlated both within and across domains. This concurrence implies that most risks are concentrated among a small group of people, which furthermore has predominantly vulnerable socioeconomic characteristics. These findings suggest that supportive policies that tackle multiple weaknesses at the same time are particularly valuable. We discuss how our risk estimates and identified risk group characteristics can be used to target prevention policies.

This paper leaves various avenues open for future research. First, one could extend the set of shock definitions. For instance, defining shocks at the (intra-)household level or in

different domains would extend our view on risk concurrence. Second, one could investigate the stationarity of the prediction models. Our models are trained using data from a period where shock prevalences are stable over time, but it is unclear whether their predictions remain accurate under large macroeconomic shocks such as the COVID-19 pandemic or financial crises. Third, it would be worthwhile to examine which sparse sets of variables achieve reasonable levels of predictive power. We have only considered extreme cases where there is either a lot of information on individuals or very little. The authors of this paper intend to continue by studying whether the recovery from adverse events is also predictable ex-ante. Combining that with the insights of this paper would give a more complete picture of individual resilience. Supportive policy should ideally be geared towards individuals who are simultaneously at risk of facing setbacks and at risk of not recovering.

References

Adda, J., Banks, J., and Von Gaudecker, H.-M. (2009). The Impact of Income Shocks on Health: Evidence from Cohort Data. *Journal of the European Economic Association*, 7(6):1361–1399.

Broten, N., Dworsky, M., and Powell, D. (2022). Do temporary workers experience additional employment and earnings risk after workplace injuries? *Journal of Public Economics*, 209:104628.

De Klerk, M., Eggink, E., Plaisier, I., and Sadiraj, K. (2023). Zicht op zorgen. Technical report, Sociaal en Cultureel Planbureau.

Desiere, S., Langenbucher, K., and Struyven, L. (2019). Statistical profiling in public employment services: An international comparison. OECD Social, Employment and Migration Working Papers 224, OECD. Series: OECD Social, Employment and Migration Working Papers Volume: 224.

Dobkin, C., Finkelstein, A., Kluender, R., and Notowidigdo, M. J. (2018). The Economic Consequences of Hospital Admissions. *American Economic Review*, 108(2):308–352.

Einav, L., Finkelstein, A., Mullainathan, S., and Obermeyer, Z. (2018). Predictive modeling of U.S. health care spending in late life. *Science*, 360(6396):1462–1465.

García-Gómez, P., van Kippersluis, H., O'Donnell, O., and van Doorslaer, E. (2013). Long-Term and Spillover Effects of Health Shocks on Employment and Income. *The Journal of Human Resources*, 48(4):37.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2021). What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics? *Econometrica*, 89(5):2303–2339.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, 105(5):491–495.

Lindeboom, M., Llena-Nozal, A., and van der Klaauw, B. (2016). Health shocks, disability and work. *Labour Economics*, 43:186–200.

Lundborg, P., Nilsson, M., and Vikström, J. (2015). Heterogeneity in the impact of health shocks on labour outcomes: evidence from Swedish workers. *Oxford Economic Papers*, 67(3):715–739.

Mueller, A. and Spinnewijn, J. (2023). The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection. Technical Report w30979, National Bureau of Economic Research, Cambridge, MA.

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

Roos, A.-F., Diepstraten, M., and Douven, R. (2021). When financials get tough, life gets rough? Problematic debts and ill health. *CPB Discussion Paper*. Publisher: CPB Netherlands Bureau for Economic Policy Analysis Version Number: CPB discussion paper, 428.

Van Hoenselaar, F., Eijsink, G., and Rupert, N. (2023). Kwetsbaarheid en veerkracht van Nederlandse huishoudens. *DNB Occasional Studies*, 21(01).

A Details Shock Definitions

This overview presents additional details about the variables used in the shock definitions.

Income Throughout the paper, income is defined as the personal primary income. This includes a person's gross income from labor and business ownership. Labor income consists of one's gross salary (including the employee's and employer's contributions to social insurance premiums), bonus and remuneration for work that is not performed as an employee. This also includes wages in kind, such as the value of the private use of the employer's car. Income from business ownership consists of the reward of the self-employed for the use of their labor and business capabilities.

Social benefits The social benefits considered in shock definition *social_benefits* are unemployment, social assistance, illness/disability and other social security benefits. These are briefly discussed below.

- *Unemployment benefits*: Upon job loss, this entitles the recipient to at most 2 years of benefits, depending on the duration of the employment history. The amount in the first 2 months is 75% and later 70% of the monthly wage. In some 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. The wealth threshold is set at 7,000 euros. For couples and older people, different thresholds apply.

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

- *Disability benefits:* Employees who are considered disabled for more than 35% are

eligible to receive disability benefits. The maximum amount is 75% of the previous salary.

- *Other social security benefits:* This is a collection of other social security benefits, such as benefits for young disabled people, older and partially disabled unemployed employees, and older and partially disabled former self-employed persons.

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

Diagnosis Treatment Combination Hospitals have to register every diagnosis, treatment and corresponding costs as a so called Diagnosis Treatment Combinations.

B Machine Learning *LightGBM* Parameters

The *LightGBM* package is flexible and allows for a range of parameters to be set. Here we list our choice of 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.

Besides the package parameters there is one other aspect of performance worth mentioning, and that is the size of the train set. More data should lead to better models, but there are diminishing returns. We have access to even more individuals in our data, but are also constrained by computational time. In runs with approximately 0.5 and 1.5 million observations in the train set we observe slight improvements in the predictions but no major shifts in quality. We therefore feel that going beyond our approximately 2 million observations would not alter our results qualitatively.

Table B1: *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	200	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 $\sim 0.01\%$).
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.

C Prediction Evaluation with Classification

We collapse the continuous shock probabilities to a binary outcome by choosing, for each shock definition, a probability threshold above which we predict a shock realization and below which we predict no shock realization. Table C1 presents various performance metrics that are often used when evaluating the prediction quality in classification exercises. This evaluation method solely rewards correct outcome predictions and not correct risk estimates. A perfect estimate of the true underlying risk would still result in some wrongly predicted outcomes.

The AUC (Area Under the Curve) is the value of the area under the ROC (Receiver Operating Characteristic), which plots the true positive rate against the false positive rate across classifier thresholds. Its value can range from 0.5 (for a model with no predictive power) to 1 (for a model with perfect predictive power). The other performance metrics are calculated using the classifier threshold that maximizes the F1-score in the test set, which implies that their values are somewhat inflated. The F1-score is the harmonic mean of the Precision (number of true positives relative to total number of positives) and Recall (number of true negatives relative to total number of negatives), while the Accuracy is the number of correct predictions relative to the total number of predictions.

Table C1 shows that all twelve shock definitions achieve good values for the AUC metric, indicating that all of our prediction models have adequate discerning capabilities. The prediction performance is the best for shocks in the labor domain compared to the health domain and better for mental health shocks compared to physical health shocks. Unfortunately, all performance metrics are heavily affected by the low prevalence of the shocks and thus difficult to assess. The metrics are best suited for balanced applications where the prevalence is close to 50%, which is not the case in our scenario. This is confirmed by table C2, which shows the same metrics for the top ten risk percentiles where the prevalence of shock realizations is closer to 50%.

Table C1: Classification performance metrics for all shocks.

Shock	AUC	F1-score	Precision	Recall	Accuracy
social_benefits	0.94	0.48	0.50	0.47	0.98
health_expenditures	0.74	0.19	0.15	0.29	0.91
relative_drop_income	0.86	0.46	0.44	0.49	0.90
absolute_drop_income	0.87	0.48	0.47	0.49	0.91
problematic_debt	0.93	0.13	0.08	0.28	0.98
economic_dependence	0.93	0.47	0.45	0.49	0.96
physical_health_expenditures	0.70	0.12	0.10	0.14	0.95
physical_health_treatment	0.75	0.19	0.15	0.26	0.93
physical_health_ic	0.76	0.05	0.04	0.06	0.99
mental_health_expenditures	0.79	0.15	0.12	0.20	0.97
mental_health_treatment	0.81	0.19	0.16	0.24	0.97
mental_health_medication	0.78	0.18	0.15	0.24	0.95

Table C2: Classification performance metrics for top 10 percentiles of main shocks.

Shock	Prevalence	F1-score	Precision	Recall	Accuracy
social_benefits	18%	0.54	0.50	0.60	0.82
health_expenditures	13%	0.25	0.15	0.81	0.37

D Supplementary Analysis of Predictions

D.1 Variable Importance

The machine learning methods that we employ do not allow us to infer causal relationships between the input variables and the predicted outcomes, but each trained model reports a table that attempts to express the relative significance of all variable for the predictions. The main caveat of this variable importance is that it often wrongly attributes importance to categorical variables or when variables are correlated to each other. Therefore, it is prone to portraying a skewed ranking.

When we look at the top 25 variables with the highest importance for the main shock *social_benefits* it stands out that 23 of them are first lags of time-dependent variables. This corroborates the expectation that more recent information is more useful for making predictions. For the main shock *health_expenditures* this top 25 includes 16 first lags and 8 second and third lags coming from 4 variables. This could indicate a smaller complexity of variable interaction since a selection of few variables are highly informative. Both *social_benefits* and *health_expenditures* include only a single time-invariant variable in their top 25, which in both cases is the year of birth.

Table D1: Top variable importance (all of which are 1^{st} lags)

Importance	*social_benefits*	*health_expenditures*
1^{st}	Number of days on any job	Healthcare expenditures excl. GP registration
2^{nd}	Hours worked (full-time equivalent)	Healthcare expenditures
3^{rd}	Employer healthcare premium	Hospital care expenditures
4^{th}	Number of days on primary job	Maternity care expenditures
5^{th}	Income insurance healthcare premium	GP expenditures

Another interesting observation is that for both *social_benefits* and *health_expenditures* their top 25 variables with the highest importance include variables from the other domain, i.e. health variables for *social_benefits* and labor variables for *health_expenditures*. This is a confirmation of the risk concurrence we document in section 7.

D.2 Predictions with Limited Variables

It is relevant to policy makers whether all of our many variables are needed to make accurate predictions. A desirable alternative would be to use only easily accessible person characteristics. We test this by making predictions for the *social_benefits* shock with a model trained on just 11 characteristics. These include gender, ZIP code, date of birth, country of birth, marital status, housing status, household composition and migration status. They constitute the most rudimentary data that the Dutch government has access to through its personal records database (*Basisregistratie Personen*).

Figure D1: Realized prevalences and risk estimates for shock *social_benefits* when trained on only 11 prominent person characteristics.

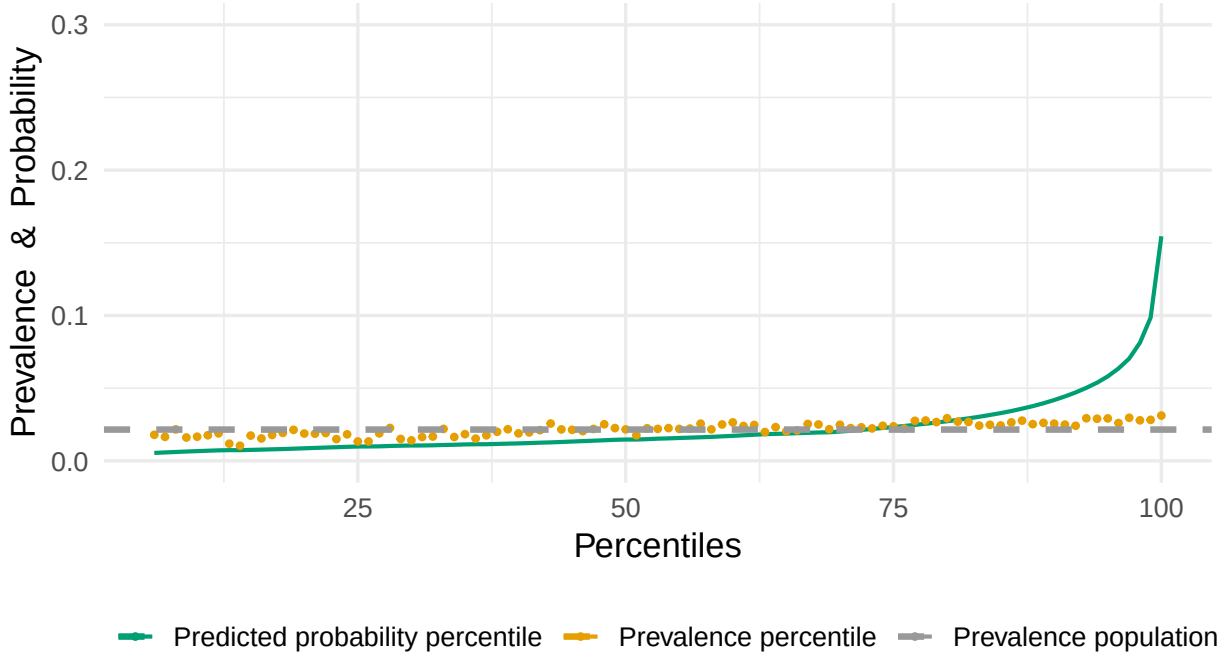


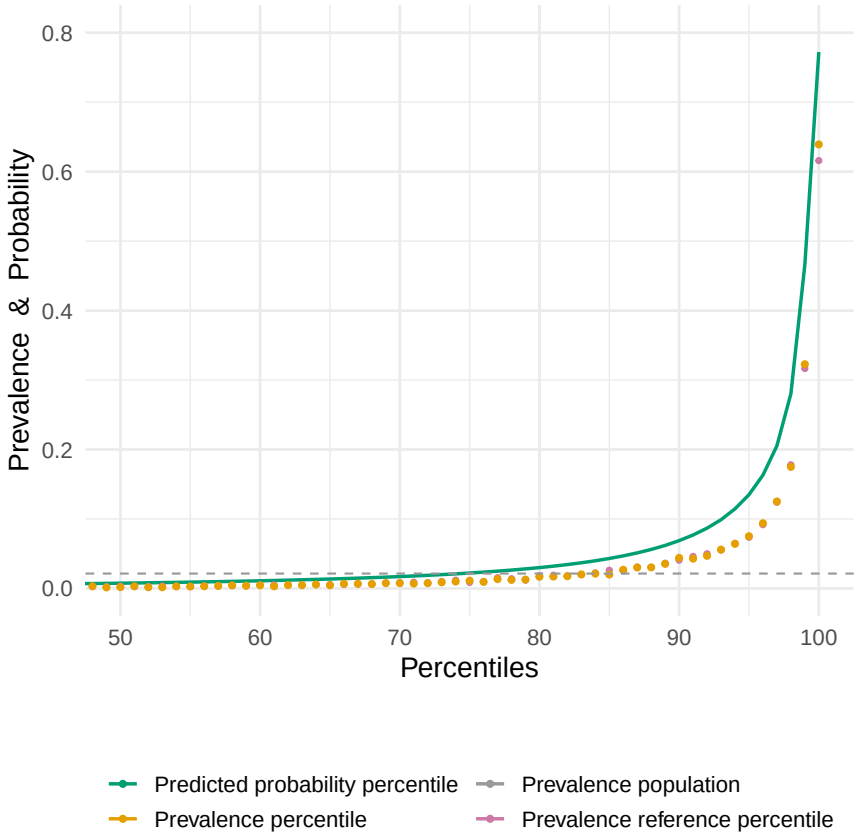
Figure D1 shows the prevalence of shock realizations per risk estimate bin, following the method from section 6. Unlike in figure 2, the realized prevalences do not trace the risk estimates at all and instead roughly follow the population prevalence. This means that model performance is extremely poor. The prediction performance found in section 5.2 cannot be reproduced with a limited variable set. Apparently, these person characteristics are not sufficient proxies for the plethora of other variables in our data set.

D.3 Predictions with Oversampling

It is common in machine learning applications to artificially increase the prevalence in the data when the sample prevalence is far below 50%. Without such balancing there is the possibility that the trained models do not optimally estimate risks because relatively few realizations were seen in the training set. We attempt to mitigate this issue through oversampling shock realizations. Because we have large amounts of data available, we do this by removing observations without shock realizations as opposed to the more common approach of adding duplicates of observations with positive outcomes. We oversample the main shock *social_benefits* to have twice the original sample prevalence.

Figure D2 shows the realized prevalences and risk estimates for each percentile bin of the risk distribution. We find that this model is better at separating risks in the upper tail of the distribution, judging from the fact that the orange dots (realized prevalences with oversampling) are higher than the purple dots (realized prevalences without oversampling) in the upper percentiles. This comes at two costs. First, the model with oversampling is less proficient in accurately estimating the risks in the bottom half of the distribution compared to the baseline model. Second, the predicted risk estimates (green line) no longer trace the realized prevalences (orange dots), because the model was trained to expect twice the prevalence that it encounters in the test set. This implies that, without rescaling, the risk estimates no longer represent probabilities.

Figure D2: Realized prevalences and risk estimates for shock *social_benefits* when trained on oversampled data with twice the original prevalence. Also shows reference prevalence from figure 2(a).



D.4 Predictions with Alternative Shock Definitions

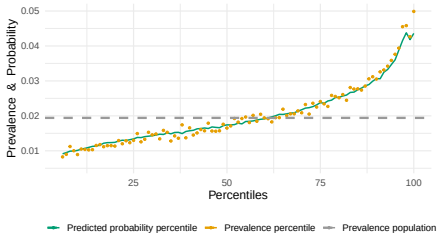
Rather than predicting the incidence of a single shock, we have also experimented with predicting the joint incidence of two shocks (either in the same period or with one shock preceding the other). The resulting prevalence of the joint shock definitions turned out to be too low in order to obtain accurate predictions despite the strong degree of risk concurrence that we find. More importantly, the joint shock definitions would yield only one set of risk estimates per combination of shocks whereas the risk estimates for specific shocks are used in section 7 to investigate how different risks relate to each other. Focusing on joint shock definitions would rule out such analysis.

We have also experimented with the time horizon of the shock definitions. Instead of requiring that shocks occur in a given year, we allowed them to occur at any point in a period of two or more years. This could improve predictive power because the trained prediction model has to be less precise about the specific moment at which a shock materializes, but it could also hurt predictive power because it is more difficult to predict what happens further in the future. We found that the predictive power of labor shocks deteriorated slightly and that it marginally improved for health shocks. This could mean that one's labor market position evolves rapidly with recent data being highly informative, while one's health is a more slow-moving object whose deterioration can manifest over a prolonged period. However, we stress that qualitatively the results were similar to our baseline findings.

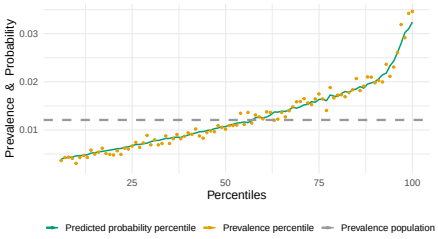
E Additional Figures and Tables

Figure E1: Joint risk distribution of *social_benefits* (on x-axis) and various health shocks (on y-axis).

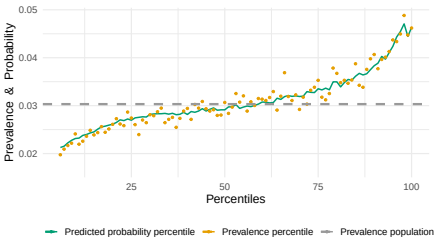
(a) *mental_health_medication*



(b) *mental_health_treatment*



(c) *physical_health_treatment*



(d) *physical_health_ic*

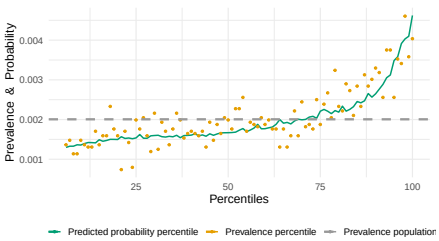
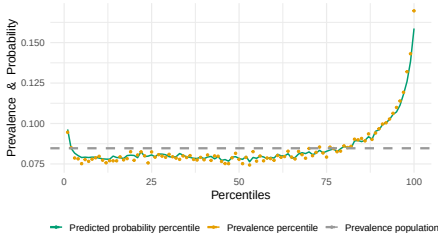
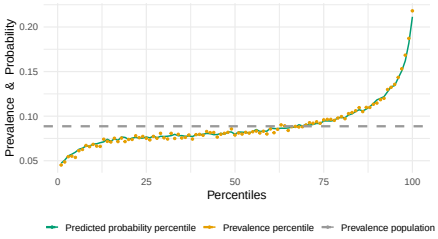


Figure E2: Joint risk distribution of *health_expenditures* (on x-axis) and various labor shocks (on y-axis).

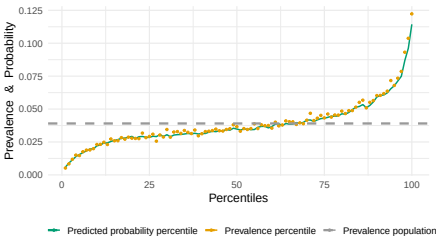
(a) *absolute_drop_income*



(b) *relative_drop_income*



(c) *economic_dependence*



(d) *problematic_debt*

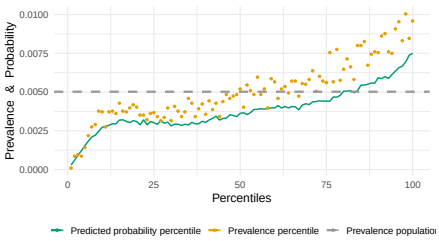
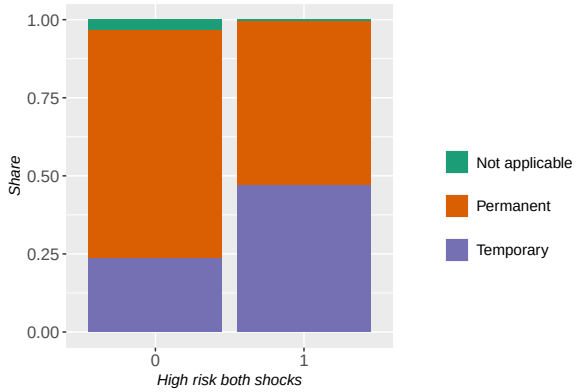


Table E1: This table displays the multiplication factor of the prevalence of a shock in year t for individuals that experienced a different shock in year $t-1$, relative to the unconditional prevalence of experiencing that shock in year t.

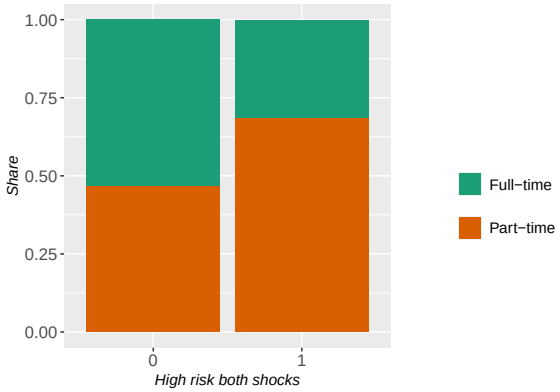
shock in t:	unconditional prevalence (%)	*conditional on shock in t-1:* *social_benefits*	*relative_drop_income*	*absolute_drop_income*	*problematic_debt*	*economic_dependence*	*health_expenditures*	*physical_health_expenditures*	*physical_health_treatment*	*physical_health_ic*	*mental_health_expenditures*	*mental_health_treatment*	*mental_health_medication*
social_benefits	2.3		4.3	3.5	4.5	4.6	2.8	1.9	1.7	3.5	3.9	3.9	3.3
relative_drop_income	8.9	4.0		2.4	2.7	3.1	1.8	1.5	1.3	1.8	2.3	2.2	2.0
absolute_drop_income	8.5	3.1	2.3		2.1	1.7	1.7	1.4	1.3	1.8	2.0	2.0	1.7
problematic_debt	0.5	3.4	2.3	1.7		2.7	1.5	1.2	0.9	2.0	2.5	2.5	2.1
economic_dependence	3.9	8.6	4.5	2.9	3.8		2.0	1.5	1.4	2.4	2.8	2.8	2.5
health_expenditures	3.6	1.5	1.1	1.0	1.5	1.2		1.4	1.9	1.8	1.8	2.1	2.2
physical_health_expenditures	2.2	1.3	1.0	1.0	1.2	1.1	1.6		2.2	3.0	1.4	1.5	1.7
physical_health_treatment	3.3	1.3	1.0	1.0	1.1	1.1	2.9	3.6		2.3	1.6	1.7	1.9
physical_health_ic	0.3	1.7	1.1	1.0	1.5	1.3	5.2	3.9	3.3		2.2	2.1	2.2
mental_health_expenditures	1.3	2.2	1.4	1.1	2.2	1.5	1.9	1.6	1.8	2.7		14.7	4.1
mental_health_treatment	1.5	2.3	1.4	1.2	2.3	1.6	2.6	1.6	1.9	3.2	10.6		4.3
mental_health_medication	2.3	2.1	1.3	1.1	2.0	1.5	3.0	1.9	2.2	3.1	5.8	5.5	1.0

Figure E3: Risk groups for the shocks *social_benefits* and *health_expenditures* jointly by employment, income and wealth characteristics. Individuals who are in the top 5% of the risk distribution of both shocks are considered at high risk for both shocks.

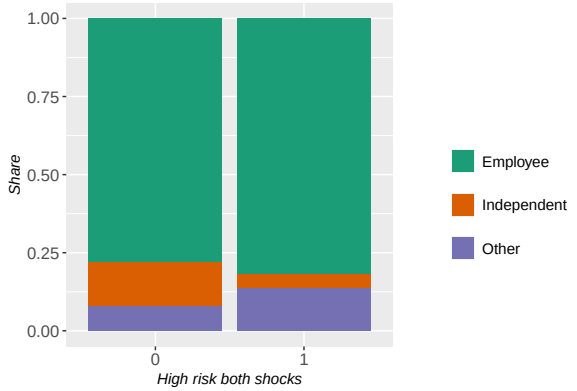
(a) Contract type



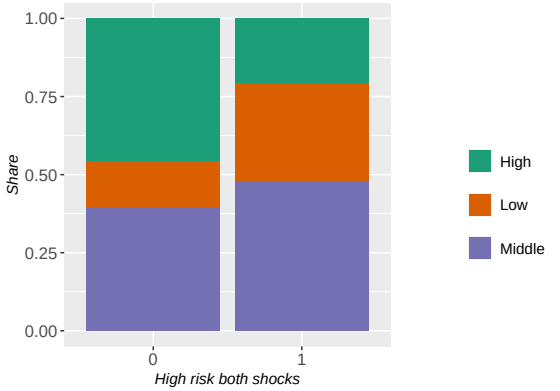
(b) Employment type



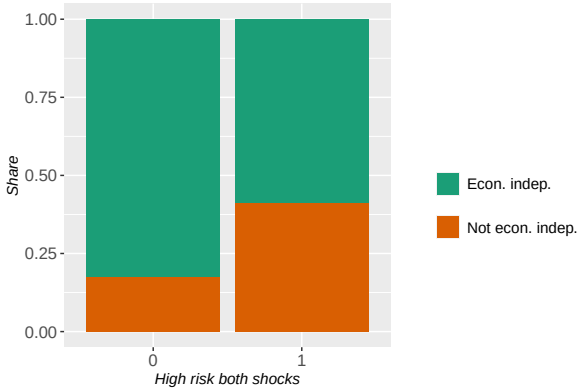
(c) Socioeconomic category



(d) Educational attainment



(e) Economic independence



(f) Wealth percentile

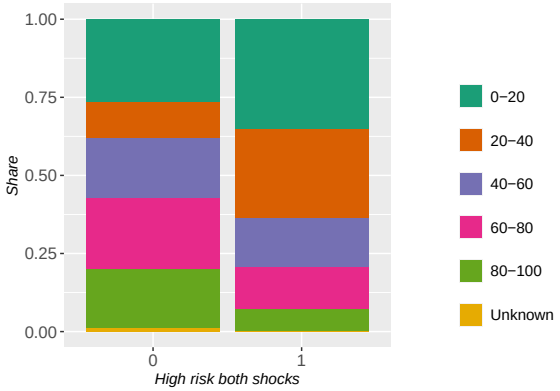
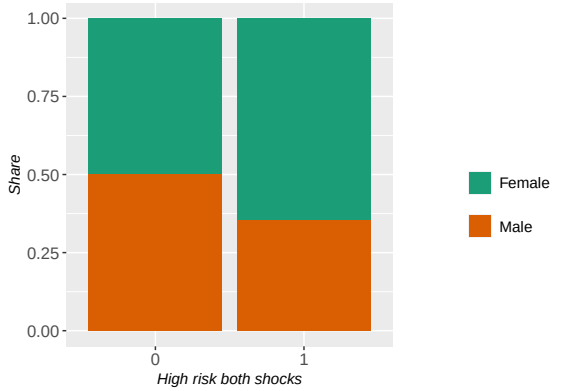
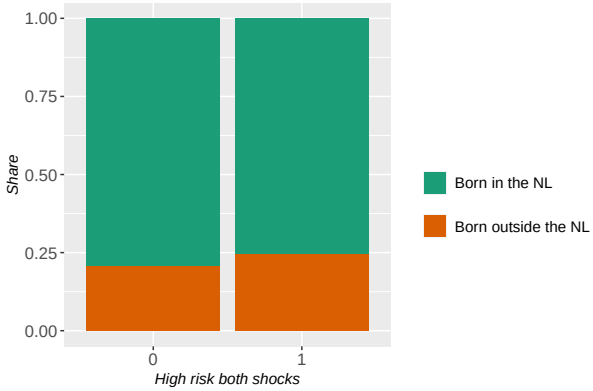


Figure E4: Risk groups for the shocks *social_benefits* and *health_expenditures* jointly by a selection of personal and household characteristics. Individuals who are in the top 5% of the risk distribution of both shocks are considered at high risk for both shocks.

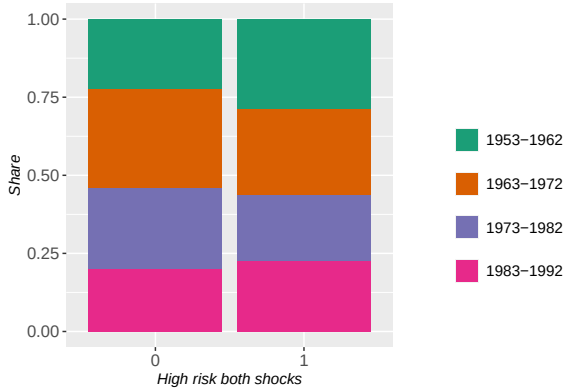
(a) Gender



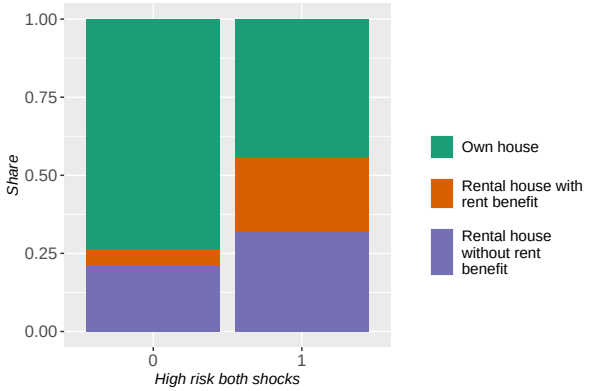
(b) Country of origin



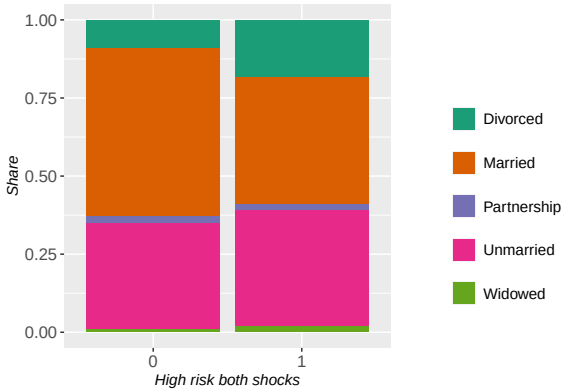
(c) Birth cohort



(d) Housing



(e) Marital status



(f) Household composition

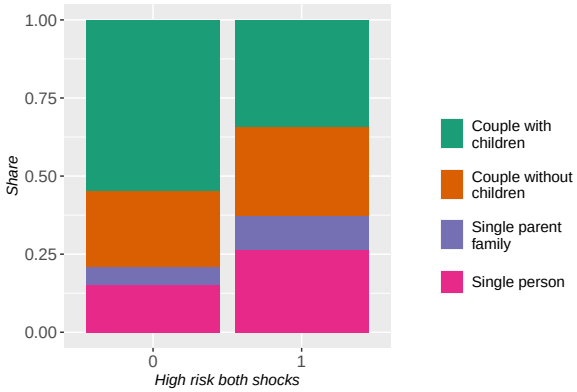


Table E2: Targeting effectiveness of *social_benefits* in various sub-populations. The first column describes selected sub-populations based on risk estimates, pre-shock socioeconomic characteristics and past shock incidence (or combinations thereof). Subsequent columns show the shock prevalence within each sub-population, the percentage of individuals within each sub-population that end up in the upper tail of the risk distribution, and the size of each sub-population relative to the total population.

	Prevalence	Tail 5%	Targeted fraction population
total population	2.3%	5.0%	100.0%
top risk distribution			
top 5% risk	29.6%	100.0%	5.0%
top 2% risk	51.5%	100.0%	2.0%
top 1% risk	67.0%	100.0%	1.0%
socioeconomic characteristics			
female	2.5%	5.6%	49.9%
rental house with rent benefit	7.6%	22.4%	5.6%
not economically independent	4.5%	11.8%	17.7%
lower education level	5.5%	14.6%	9.5%
single	3.2%	7.8%	15.3%
temporary contract	4.9%	12.5%	20.0%
born outside the NL	3.5%	8.4%	20.6%
2nd wealth quintile	4.5%	11.9%	11.9%
shock realization in t-1			
relative_drop_income	9.9%	19.8%	5.9%
absolute_drop_income	7.9%	14.9%	5.8%
problematic_debt	10.2%	23.3%	0.4%
economic_dependence	10.5%	21.9%	1.8%
health_expenditures	6.4%	14.0%	2.3%
physical_health_expenditures	4.4%	9.7%	1.6%
physical_health_treatment	3.9%	9.0%	1.9%
physical_health_ic	7.9%	17.2%	0.2%
mental_health_expenditures	9.0%	20.0%	0.9%
mental_health_treatment	8.9%	20.4%	1.3%
mental_health_medication	7.6%	15.7%	1.7%
combinations of socioeconomic characteristics and shock realizations in t-1			
top 2 socioeconomic characteristics	8.7%	31.8%	2.6%
top 3 socioeconomic characteristics	14.3%	45.5%	0.5%
top labor shock and socioeconomic characteristic	18.7%	41.0%	0.1%
top health shock and socioeconomic characteristic	20.5%	47.5%	0.3%

Table E3: Targeting effectiveness of *health_expenditures* in various sub-populations. The first column describes selected sub-populations based on risk estimates, pre-shock socioeconomic characteristics and past shock incidence (or combinations thereof). Subsequent columns show the shock prevalence within each sub-population, the percentage of individuals within each sub-population that end up in the upper tail of the risk distribution, and the size of each sub-population relative to the total population.

	Prevalence	Tail 5%	Targeted fraction population
total population	3.6%	5.0%	100.0%
top risk distribution			
top 5% risk	16.5%	100.0%	5.0%
top 2% risk	21.6%	100.0%	2.0%
top 1% risk	25.6%	100.0%	1.0%
socioeconomic characteristics			
female	4.2%	6.4%	50.6%
rental house with rent benefit	5.8%	13.0%	10.6%
not economically independent	5.4%	11.7%	27.6%
lower education level	5.1%	9.8%	13.0%
single	4.4%	8.5%	17.3%
temporary contract	3.2%	3.3%	19.3%
born outside the NL	3.7%	4.7%	22.9%
2nd wealth quintile	5.0%	9.9%	15.7%
shock realization in t-1			
social_benefits	5.6%	12.4%	2.2%
relative_drop_income	4.0%	6.6%	7.5%
absolute_drop_income	3.7%	5.4%	6.7%
problematic_debt	5.6%	9.0%	0.6%
economic_dependence	4.4%	7.8%	2.9%
physical_health_expenditures	5.2%	7.6%	1.7%
physical_health_treatment	7.0%	13.8%	2.1%
physical_health_ic	6.7%	34.8%	0.2%
mental_health_expenditures	6.5%	13.6%	1.2%
mental_health_treatment	7.6%	15.8%	1.7%
mental_health_medication	7.9%	16.2%	2.1%
combinations of socioeconomic characteristics and shock realizations in t-1			
top 2 socioeconomic characteristics	6.4%	5.4%	2.4%
top 3 socioeconomic characteristics	6.5%	6.3%	0.7%
top labor shock and socioeconomic characteristic	6.4%	13.0%	0.2%
top health shock and socioeconomic characteristic	9.7%	22.3%	0.5%