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]

December 14, 2023

Abstract

This paper employs machine learning techniques to estimate the ex-ante probabilities that individuals will be confronted with adverse labor and health events. Using rich administrative data on the entire Dutch population, we document that shock incidence is predictable especially in the labor domain and to a lesser degree in the domain of mental and physical health. Moving from ex-post incidence to ex-ante probabilities allows us to separate predictable components of shocks, interpreted as ex-ante risk types, from random components. We bring together risk estimates for many different shocks and uncover that risk concurrence is sizable, monotone, non-linear and extended across domains. We show that socioeconomic characteristics pertaining to employment status, educational attainment, migration background, income and wealth are over-represented in the upper tails of the estimated risk distributions. Lastly, we discuss the implications of our findings for targeted prevention policies based on individual risk estimates, shock (co-)incidence and person characteristics.

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). An important, yet unexplored, question is whether these shocks occur randomly, or whether they carry an element of predictability that could be acted upon. Exploring this predictable element not only unveils the distribution of this predictable risk factor across the population, but also sheds lights on the correlation between the probabilities of encountering shocks across different domains. This paper brings together data on millions of people to study the predictability and (co-)incidence of a wide array of labor and health shocks, and discusses what this could mean for social insurance policies and targeted prevention policies in particular.

The literature on shocks has so far focused on the realizations of adverse events. However, the materialization of a shock can be seen as the combination of a predictable and a random component. Following Mueller and Spinnewijn (2023), we interpret the predictable component based on observable characteristics as an individual's ex-ante risk *type* and employ machine learning to estimate it.[1] Rather than analyzing ex-post realizations, our research uncovers the entire distribution of ex-ante probabilities of adverse events. In addition, we do not only look at shocks in isolation, but also at how these risk types are distributed across different domains. By doing so, the narrow study of shock coincidence becomes the richer study of risk concurrence, which constitutes a novel contribution to the existing literature.

We use machine learning techniques to estimate the probability that individuals are confronted with labor and health shocks. We deploy machine learning methods for three reasons. Firstly, our aim is to unveil ex-ante risk types by disentangling the predictable and random risk elements. Leveraging the aptitude of machine learning for classification

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

problems, we obtain accurate estimations of the predictable risk component. Secondly, machine learning is capable of dealing with large amounts of data. We collect administrative data encompassing labor, health, socioeconomic and demographic information for the entire Dutch population from 2013-2018, comprising a total of over five hundred variables observed at yearly frequency for over twenty million people. Thirdly, machine learning can discern 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 twelve shocks that we consider across the labor and health domains.

Our paper shows that the incidence of shocks can be accurately predicted as indicated by conventional metrics for assessing prediction performance.[2] This is especially the case in the labor domain and to a lesser degree in the domain of mental and physical health. After dividing the risk distribution in percentile bins, we compare the average predicted probability in each bin to the actual shock prevalence. It turns out that the group-level predictions and realizations line up very closely throughout the entire distribution.

The estimates 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. However, this large amount of 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.

Next, 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

[2]We only use past information of the individual (with a lag of up to three years) to predict future outcomes.

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 association 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.

We then dive into the characteristics of the people in the upper tail of the risk distributions (both for singular shocks and for combinations of two) and find that certain subgroups are strongly and persistently over-represented. Notable among these groups are individuals 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 firmly underscores the unequal distribution of risks within the domains of labor and health across the Dutch population.

The finding that risk types are estimable, unequally distributed and extended across domains, is relevant to policymakers. Social insurance policies play a major role in protecting people against setbacks. However, existing policies tend to fall short by responding reactively 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 pro-active policy could break up chains of adverse events. Preventive measures, such as improving the labor market position of individuals at high risk of job loss, could be more effective, beneficial for well-being, and cost-efficient than addressing the aftermath of setbacks. 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.

Our prediction models can accurately pinpoint groups of individuals that bear the most risk. While we use machine learning techniques primarily to study risk concurrence, one could wonder if the trained models can also be used to target prevention policies on an individual level. It is unclear whether it is currently feasible or desirable to do so, especially because we find that sufficient prediction performance is only achieved after assembling many variables spanning multiple domains in a timely manner. We show that targeting based on both readily observable person characteristics and past realizations of shocks is a reasonable alternative. It is already common in practice to rely on person characteristics, but we find that past realizations of other setbacks are equally, if not more, informative indicators of risk exposure. This follows directly from the pervasive concurrence of risks that we document.

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 evaluate the machine learning models. Section 4 presents the shock definitions in the labor and health domain, and the rationale behind them. Section 5 explains the methodology employed to get to the risk estimates of individual shocks and evaluates the prediction quality. Section 6 describes how the risk distributions of different shocks relate to each other. Section 7 explores the characteristics of individuals in the upper tail of the risk distributions. Section 8 discusses the policy implications of our results. Section 9 concludes.

2. Related Literature

In the following section we contextualize our study within the existing literature concerning labor and health shocks, and we discuss some relevant applications of machine learning.

Relation between labor and health shocks. An extensive literature has concerned itself with the link between labor and health shocks. Adda et al. (2009) focus on the effect of income shocks on health over the life cycle. They exploit a number of exogenous changes in income and find that income shocks affect health behavior and mortality. A study by Guvenen et al.

(2021) on US data shows that negative income shocks are concentrated in a small group and that the individual risk of such a shock strongly depends on one's income and life course. This study also reports that individuals in the group experiencing the largest income shock have disproportionately often become occupationally disabled during that period, even when correcting for disability insurance income. The study by Roos et al. (2021) reports that on average the costs of mental health care rise by 30% after people encounter financial problems. Moreover, the likelihood of using mental health care services, as well as social or financial assistance, increases too.

The reverse direction of the causal effect has also been extensively studied. García-Gómez et al. (2013) investigate the causal effect of unexpected hospitalizations on labor income in the Netherlands, utilizing administrative data from Statistics Netherlands. Similar inquiries have been conducted for Sweden and England, as evidenced by Lundborg et al. (2015) and Lindeboom et al. (2016), respectively. The common thread in this literature underscores that occurrences of health shocks exert a prolonged, adverse influence on labor income. Moreover, the effect appears to be contingent upon the form of employment contract, see Broten et al. (2022). In the United States, Dobkin et al. (2018) have examined how health insurance mitigates this detrimental effect.

While our paper refrains from drawing causal inferences, it is worth noting that the well-documented bidirectional relationship between labor and health shocks in the existing literature is also evident in our findings. Our analysis demonstrates that, given the occurrence of a shock in one domain, the likelihood of experiencing a shock in the other domain increases significantly.

Machine learning. The aforementioned strand of literature on labor and health shocks generally takes a limited perspective on income and health risks. To establish a causal relationship, the shock must meet the exogeneity condition. 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 unex-

pected hospitalization versus those who did not. However, Kleinberg et al. (2015) argue that there are many policy applications, so called *prediction policy problems*, where causal inference is not central, or even necessary. Instead, accurate prediction can generate large impact. This is also the approach of this paper - rather than establishing a causal relationship, we aim to predict the (co)incidence of adverse shocks. Therefore, we are not restricted to analyzing shocks that meet the exogeneity condition.

The emerging trend in the literature on prediction policy problems emphasizes how machine learning can be used to assess individual risks and how these estimates can be applied to enhance the efficiency of policies. The added value of machine learning over traditional econometric methods lies in the fact that machine learning models have much stronger predictive power due to their flexibility in incorporating complex interactions and large numbers of variables. Mullainathan and Spiess (2017) provide an extensive overview of machine learning applications in economic prediction problems and policy domains.

The OECD study by Desiere et al. (2019) argues that the increase in available data and computing power has made statistical profiling, i.e. the use of statistical models to predict an outcome, more common. The study gives an overview of how different countries use statistical profiling to predict long-term unemployment and how these predictions are translated into policy actions. In the realm of academic research, Mueller and Spinnewijn (2023) study the predictability of long-term unemployment using Swedish administrative data. Compared to using standard socio-demographic information, adding data on job seekers' employment history prior to becoming unemployed significantly enhances the model's predictive power. We contribute to this line of research by exploring more shocks within the labor domain, and also including shocks in the health domain.

Athey (2017) underscores that there is a gap between making a prediction and making a decision, and that understanding underlying assumptions is necessary in order to optimize data-drive decision-making. For example, pure prediction methods do not answer the more complex question of which individuals respond best to a policy intervention. Additionally,

we argue that data availability is another factor that creates a delay between prediction and policy action.

In the paper by Einav et al. (2018), machine learning techniques are employed to estimate mortality risks. Instead of counting instances of death, the researchers apply machine learning methods to estimate ex-ante mortality probabilities. By predicting probabilities instead of focusing on outcomes, individual-level mortality risks can be determined. They conclude that predicting short-term mortality is feasible only for a very small group. In our study, we take a similar approach by focusing on ex-ante probabilities, rather than ex-post realizations. In contrast to the aforementioned study, we find that it is possible to identify a substantial group of individuals with a high risk of experiencing a labor or health shock.

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 show significant overlap with those groups we see over-represented in the upper tails of our estimated risk distributions. By focusing on estimating ex-ante probabilities rather than ex-post realizations, our study distinguishes itself from much of the existing literature on the vulnerability of individuals to adverse events. The ex-ante perspective 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 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 in the three aforementioned domains together span roughly five hundred variables and 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.[3] 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 less complete.[4]

In each year, roughly seventeen million individuals are present in the data set, but we shrink the sample in various ways. Firstly, we exclude individuals aged below 25 years or above 60 years. Since our aim is to study (the interaction of) generic 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

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

[4]Unfortunately, it is not possible for us to assess how the lower health data coverage influences our results.

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.

probable that labor and health risks are particularly related as people approach retirement due to 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. Secondly, we exclude individuals with an unknown or uncommon household composition such as student housing or care homes. Thirdly, 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. Our aim is to characterize the risk distribution of the general population that is well-covered by the national statistical office.[5] 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. 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 the models' 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, making it relevant for policymakers to act on. Additionally, the machine learning model has good predictive power for shocks with this prevalence (see section 5). 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

[5]High-risk individuals that are not present in administrative records are likely not covered by social insurance policies and the government has different polices to assist them.

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 0 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 understood as the occurrence of a particular event, such as hospitalization or becoming a recipient of social benefits. Alternatively, a shock can manifest itself more 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, albeit not impossible in some cases. For example, there exists some flexibility in deciding which types of social benefits to include in the shock definition *social_benefits*. 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 ways (see table 2).

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. With an eye on the possibility of using this study to explore the potential of prevention policy, we focus on new adverse situations. 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.[6]

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 $t-1$

[6]We briefly return to this point in section C.4.

cannot receive the shock *absolute_drop_income* in year t by construction. Consequently, our prediction model assigns a near-zero risk estimate to the individual-year observations that do not meet these preconditions (for more details on the shock predictions, see section 5). 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.[7]

For the sake of simplicity, most of the analysis concentrates on two key 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 key 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.[8] This shock has a prevalence of 2.3% in the sample.

The key 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 sample 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

[7] In section 6 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%.

[8] 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 does no longer stem from their employment or business.

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 shock represents a deliberate choice rather than a setback. 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. Predictability of Shock Incidence and Risks

In this section we employ machine learning methods to move from shock realizations to shock probabilities (i.e. risk estimates or types). In subsequent sections we use the risk estimates obtained from these predictions as the basis for further analysis. We show that we are able to make predictions that accurately reflect the ex-ante chance of someone facing a shock, and that we can pinpoint groups with significantly higher risks.

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. Using this set of binary indicator variables $s_{i,t}$ for all individuals i and years t in our data, we can perform basic analyses on characteristics of people that end up experiencing shocks and of those that do not.

We can, however, extend our data to allow for more extensive and insightful analyses. The observations of shocks that we have are limited to binary realizations. What this data fails to accurately capture is the underlying probability that individuals had of facing a shock realization. This probability is what we are actually interested in, since it tells us how much at risk different individuals are. From a mere realization it is impossible to tell whether an individual was very unlikely to experience a shock, but was simply unlucky enough to

do so anyway, or whether an individual was very likely to experience a shock, but was lucky enough to avoid doing so this time. When we discuss targeted prevention policies and relations between setbacks, it is these underlying probabilities that we are most interested in. Instead of identifying which individuals did and did not experience a shock, we would prefer to identify which individuals were at a high or low risk to do so.

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 do this by having a machine learning model predict shock realizations for individuals based on their data from previous years. Instead of a binary categorization task (1 for shock, 0 for no shock) we can ask the machine learning model to provide a score for each category (because the categories are 0 (no shock) and 1 (shock), this is a number between 0 and 1). If calibrated, we may then interpret this score as representing a risk estimate: the probability that the outcome will be a shock realization (as opposed to no shock realization). In principle, these risk estimates can be taken to be the underlying probabilities $p_{i,t}$ of an individual i having a shock realization in year t.

For us to be able to use the computed shock risk estimates as if they were the actual underlying shock probabilities we must be assured that our machine learning predictions are sufficiently accurate. This is also a result in itself, as it answers the question *"Can individual shocks be accurately predicted from extensive personal data?"*. Fortunately, our data and method allows us to verify the prediction quality. When we talk about predictions, we mean predicting past events from data prior to that event. In that sense the predictions are true agnostic predictions. However, in our data we can see the actual realizations of those events that the machine learning model is agnostic of. Hence, we can evaluate the accuracy of the predictions made by the machine learning model as compared to the observed realizations. More challenging to evaluate is the quality of the risk estimates obtained from the predictions, since the true underlying shock probabilities (the real-world data generating process) cannot

be observed. Nevertheless, we are able to draw convincing conclusions about prediction quality regarding risk estimates as well.

Important to note is the way in which we can employ the risk estimates $p_{i,t}$. Our aim is to open up new and more complete avenues for analyses of which individuals (are likely to) experience shocks. The risk estimates $p_{i,t}$ allow for an extended continuous representation of shock data, as opposed to the limited discrete binary representation offered by the indicators $s_{i,t}$. What we cannot do, however, is extract information directly from the model that was used to make the predictions since the methods we employ do not allow us to infer direct causal relationships between the input variables and the predicted outcomes. Thankfully, our analysis does not require a causal interpretation of how the model arrives at its risk estimates. Once we validate that the estimated risks align with the realizations we observe in the data on average, we may treat those risk estimates as reasonable representations of the latent risk exposure to setbacks and subsequently perform analysis based on them.

5.1. Machine Learning Method

To make predictions, we use the *R* package *LightGBM*.[9] This package implements gradient boosting machine learning methods, i.e. it builds a prediction model as an ensemble of simpler models (see Friedman (2001)). The simple models that we train are decision trees, which means that our chosen method is gradient boosted trees. We train our gradient boosted trees model 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.

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.[10] For each individual, we have both time-invariant and time-dependent

[9] For the full set of parameters that were used when training the *LightGBM* models, see appendix B.

[10] 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.

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 make sure we are making purely agnostic predictions, i.e. to avoid data leakage, an observation includes no time-dependent information relating to the year for which the prediction is made. A time-dependent variable relating to the year of the observation (and thus the outcome), could contain information about whether the shock occurred. Including this information in our model would therefore lead to spurious predictions. Of course, each observation does have a shock outcome in the year of the observation associated with it. In the train set this outcome is what is used to supervise the training, while in the test set this outcome is the objective of prediction.

For each shock definition, we train one single model to make predictions for all observations in the test set. This means that we do not train on data from a certain year to make predictions for that year. The train and test sets contain observations from across all years that we have data for. Note that although we do not train and predict for each year separately, the machine learning method has access to the year variable and could thus learn year fixed effects and apply them accordingly in its predictions.[11] 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. No individual shows up in both the train and test set.

5.2. Evaluating Quality with Shock Predictions

Our analysis requires that the shock predictions and risk estimates that we obtain from the trained models are of sufficient quality. In this section we evaluate shock prediction quality using statistics common to machine learning applications, such as the F1-score and

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

the AUC. Both of these evaluate the ability of the model to discern between outcomes (shock realization *vs.* no shock realization). The F1-score combines the precision and recall by taking their harmonic mean. Precision expresses how well Type 1 errors are avoided, and recall expresses how well Type 2 errors are avoided. The AUC (Area Under the Curve, or c-statistic) 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. The AUC aggregates the plot of these two rates into a number, with 1 representing a perfectly discerning model. These common statistics are suited for point-prediction evaluation, but are unable to directly evaluate risk estimates. Because the latter is more important to us, we use another method to also explicitly evaluate the quality of risk estimates in section 5.3.

To calculate the mentioned statistics we need to collapse our risk estimates into point-predictions. We thus need to decide when we have predicted a positive outcome. This requires choosing a classifier threshold for the risk estimates above which we say that we have predicted a positive outcome. The model is evaluated along the optimal threshold that best reveals its capability to discern between outcomes. Because we choose this optimal threshold in the test set, our statistics are on the optimistic side. Since we know the outcomes for the test set, we can evaluate the predictions that we obtain by comparing them to the outcomes. With the optimal threshold and the real outcomes each prediction is now categorized as either truly or falsely having predicted either a positive or negative outcome. This allows us to compute the precision and recall and with them the F1-score and AUC, shown in table 3.

We see that predictions of shock *social_benefits* have a high AUC, with predictions of shock *health_expenditures* being somewhat less discerning but still with a respectable AUC. However, we immediately see that the evaluation of the discerning capability is heavily affected by the low shock prevalences. For example, the accuracy, which is simply the fraction of correct predictions irrespective of positives and negatives, is high simply by grace that predicting a negative outcome often results in success. Therefore, the accuracy is not well suited to evaluate our predictions. The precision and recall, as well as the F1-score

Table 3: Prediction performance metrics for all shocks.

Shock	AUC	F1	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

that combines them, provide a more insightful perspective on the relative success of our predictions and are therefore preferred over simple statistics such as accuracy. Nevertheless, the F1-score is also affected by the low prevalence to some extent: its value would be higher in a balanced setting. We can see this in table 4, which shows the same statistics for the top ten percentiles in which the prevalence of shock realizations is much higher.[12]

Table 4: Prediction performance metrics for top 10 percentiles of main shocks.

Shock	Prevalence	F1	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

The F1-score is just one way of weighting the precision and recall. Depending on policy goals, arguments can be made to reduce either Type 1 or Type 2 errors. For example, if treatment is costly, then reducing Type 1 errors would be preferred. In the model evaluation this preference can be taken into account by increasing, for this example, the weight of precision in the F-score.

[12]The accuracy drops as well, confirming that it is not a useful evaluation statistic in table table 3.

In conclusion, we can collapse the risk estimates to binary outcome predictions to obtain values for both the AUC and F1-score. These indicate that our models are able to discern next-year shock realizations. The low prevalence of shocks requires slightly adjusted interpretation of common statistics.

5.3. Evaluating Quality with Risk Estimate Percentiles

In the previous subsection we evaluated prediction quality by comparing a prediction to an outcome for individual observations. By doing this, however, we have omitted a crucial part of the prediction from the quality evaluation. Since we already have the outcomes, there is no sense in predicting them outright. Instead, we set out to predict the probability of a positive outcome, i.e. the estimated risk of suffering a shock. Using a threshold to obtain binary predictions destroys most of the information that was captured by this risk estimate.

Since we do not know the real probabilities underlying the outcomes that we observe, we cannot directly compare our risk estimates to them. We can, however, compare our risk estimates to aggregate outcomes. To do this we divide our test set into percentile bins of relative risk estimates. For each bin we know the average risk estimate. We also know the prevalence of shocks among the approximately twenty thousand individuals in each bin. If our prediction quality is good we should see that the prevalence and average risk estimate are similar in each percentile bin. The key realization in the evaluation method of this section is that a perfect estimate of the true risk would still result in some wrongly predicted outcomes due to the presence of ex-post randomness. This is inadequately captured by the evaluation metrics in section 5.2, which focus solely on rewarding correct outcome predictions, but it is captured in this evaluation method.

Fig. 1. Realization prevalence and risk estimates for shock *social_benefits*.

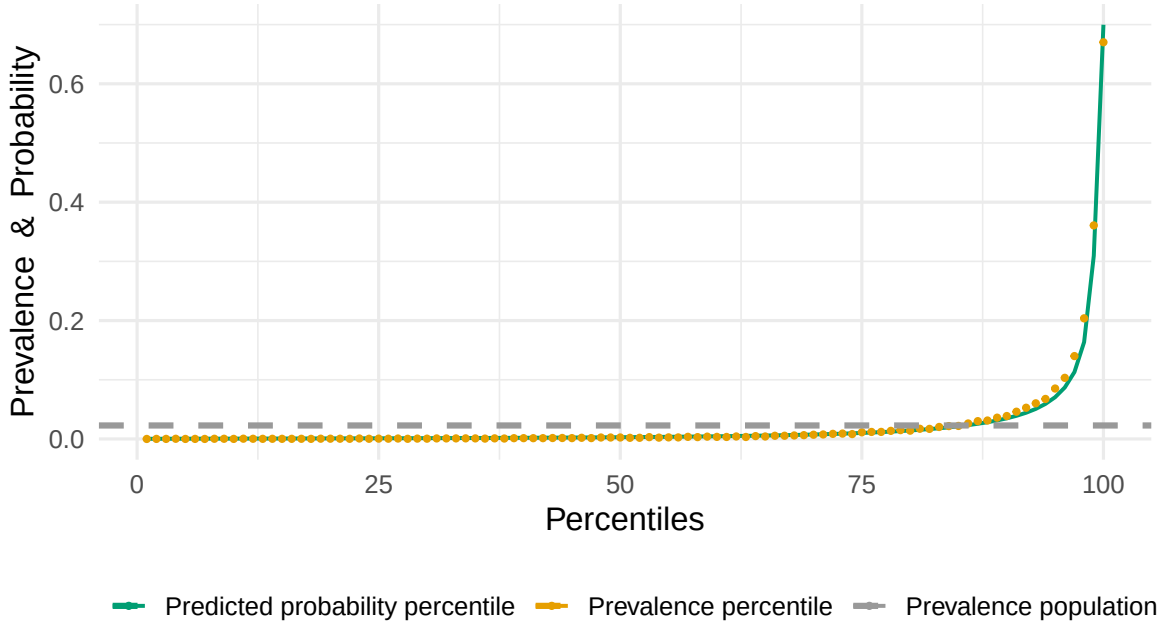
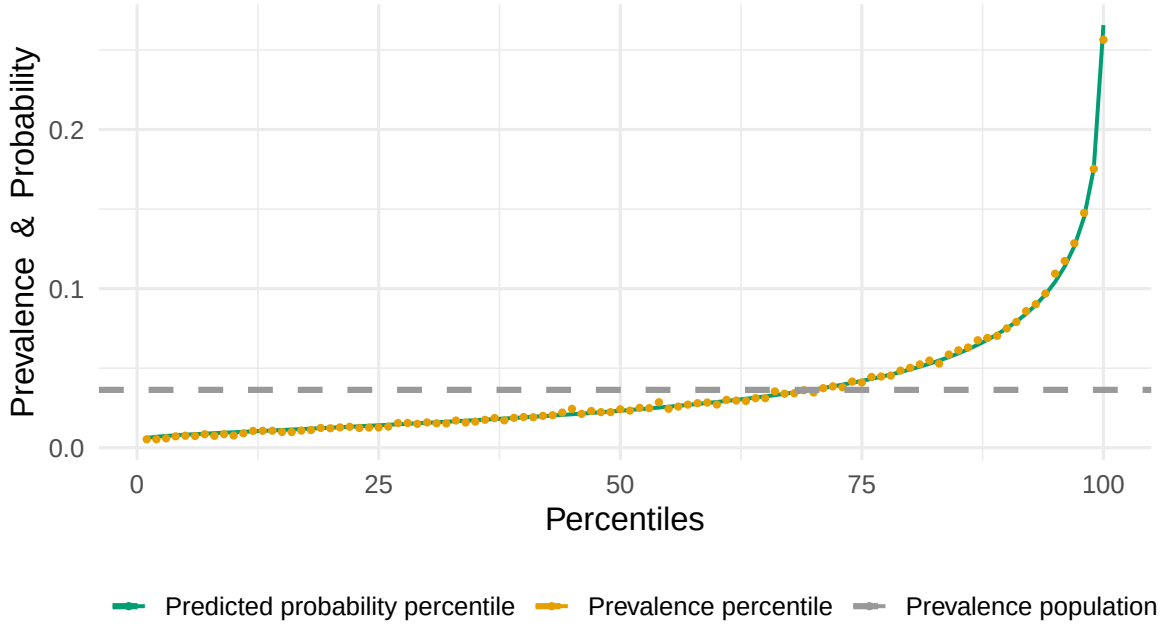


Fig. 2. Realization prevalence and risk estimates for shock *health_expenditures*.



Figures 1 and 2 show both the average risk estimate (green line) and shock prevalence (yellow dots) for each risk estimate bin. As can be seen, the risk estimates trace the prevalences almost exactly across the entire range. Evidently, we obtain accurate predictions for

both high-risk and low-risk individuals. Moreover, we see that our model is able to discern groups of individuals that bear the most risk, as evidenced by the sharp increase in both the risk estimate and prevalence towards the upper tail. The vertical differential is an indicator of this discerning ability, and in this case tells us that risk estimates for *social_benefits* are more discerning than they are for *health_expenditures*. Compared to the population average (grey stripes), the highest risk percentile faces a risk that is up to thirty (in the case of *social_benefits*) and ten times larger (in the case of *health_expenditures*). Our model is thus able to pick out a large tail of substantially heightened risk of setbacks on which it accurately matches the observed prevalences. Furthermore, over two thirds of the population face below average levels of risk. We discuss the policy implications of this finding in section 8.

5.4. Supplementary Analysis

This section thus far has presented our baseline results regarding the predictability of shock incidence. However, we have also experimented along various dimensions to gain additional insights about this predictability. These analyses can be found in appendix C. In section C.1 we take a cautious look at variable importance, in section C.2 we discuss the predictions from a machine learning model trained on a data set with only a few rudimentary variables, in section C.3 we discuss the predictions that arise when we oversample, i.e. artificially increase the prevalence of shock realizations in our train sample, and in section C.4 we discuss alternative shock definitions.

6. Risk Concurrence

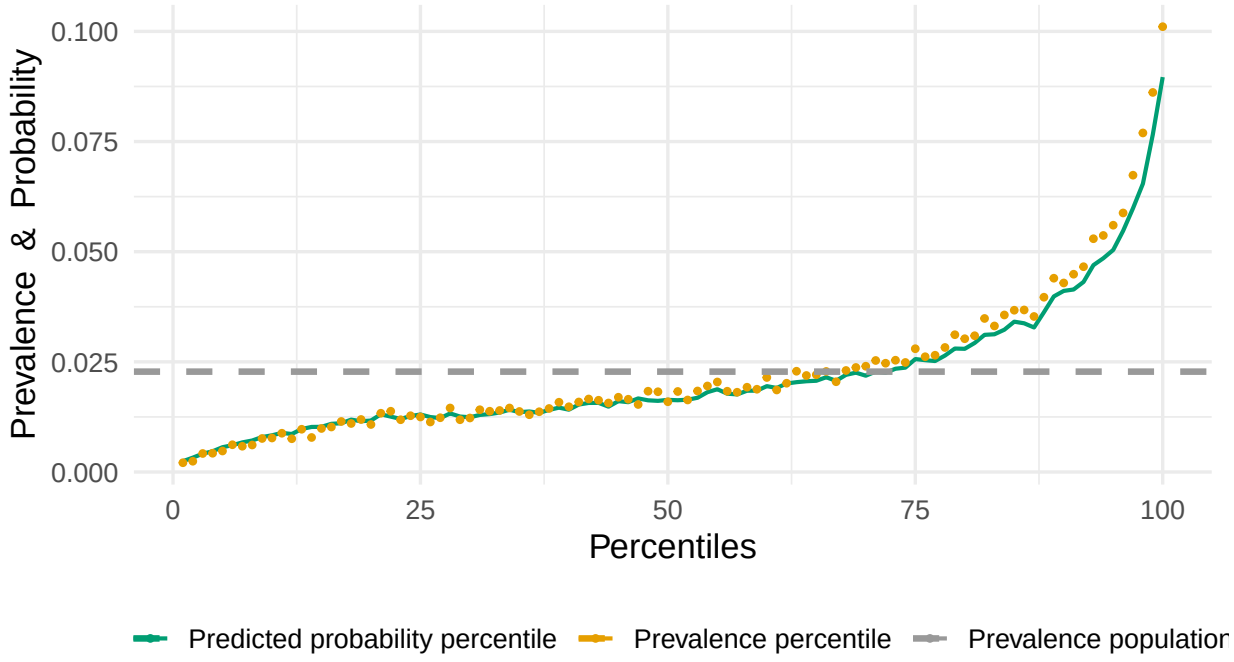
Having characterized the risk distributions of individual shocks, we now move on to study how the risk distributions of different shocks relate to each other. We will show that there is strong concurrence between risks not only within a particular domain, but also across domains. Individuals with a high 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.

6.1. Risk Distributions across Domains

The figures of section 5.3 can be amended to contrast the risk distributions across domains. For example, fig. 3 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 (yellow dots). Our machine learning model is thus not only able to identify who is most likely to be hit by an adverse event (as established section 5.3), but also able to identify who is most likely to be confronted with the joint materialisation of adverse events.

Fig. 3. Joint risk distribution of *health_expenditures* (on x-axis) and *social_benefits* (on y-axis).

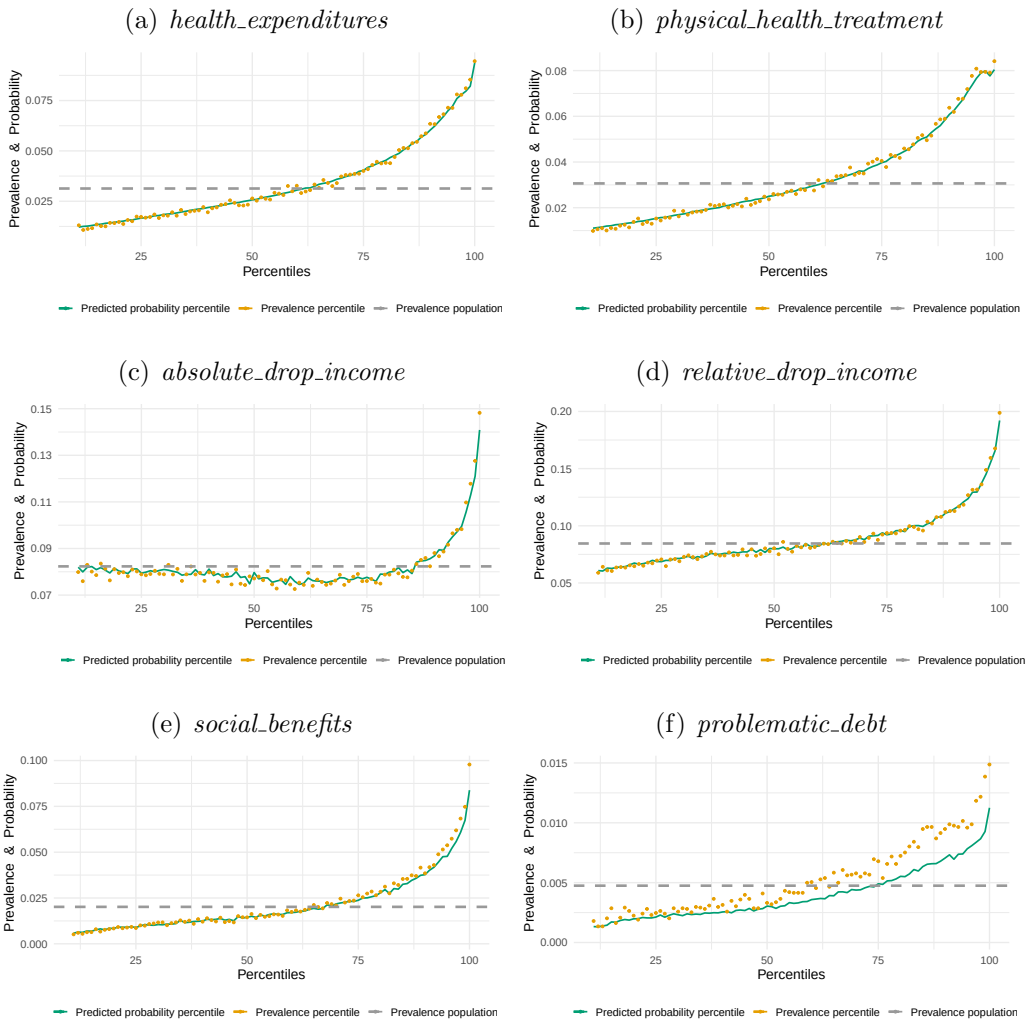


People with the lowest chance of a sharp increase in health expenditures also face the lowest probability of having to suddenly rely on social benefits, and vice versa. The inequality in risk exposure is substantial: those in the upper tail of the *health_expenditures* risk distribution are up to four times as likely to be faced with the *social_benefits* shock compared to the average population, and more than ten times as 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 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 distribution of *mental_health_medication* with various other shocks in fig. 4. 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 (fig. 4(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 (fig. 4(b)). Apparently, even within a particular domain such as health, there is

Fig. 4. Joint risk distribution of *mental_health_medication* (on x-axis) and various other shocks (on y-axis).



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 (figs. 4(c) and 4(d))[13] and in terms of having to suddenly rely on social benefits (fig. 4(e)). This extends from the labor domain into the wealth domain, where the labor income fragility

[13]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.

comes with a heightened likelihood of defaulting on mandatory health insurance premium payments (fig. 4(f)), which is a commonly used indicator for the onset of problematic debt in the Netherlands.

Taken together, the subfigures in fig. 4 illustrate 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 figs. D1 and D2, 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.

6.2. 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 5. 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 6.1 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.

Firstly, all values have a positive sign, which indicates that the positive association found in the plots of section 6.1 is also present in all other shock combinations. Secondly, 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 5 represents, we map a subset of numbers to the earlier presented joint risk distribution plots. Figure 3 amounts to a Spearman rank correlation coefficient of 0.31, fig. 4(b) to a coefficient of 0.70, and fig. 4(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 5 imply that the risk concurrence within a domain is particularly monotone.[14] 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.

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

Table 5: 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

6.3. Conditional Risk Estimates across Domains

In the introduction we shared the stylized fact that 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. One might wonder whether the risk estimates derived from our machine learning models support a similar conclusion. Table 6 depicts the factors by which the conditional probabilities are larger than the unconditional probabilities based on risk estimates, while table D1 in the appendix 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 1.0, 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 6: 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	

7. Risk Groups

Having constructed risk estimates on an individual level, a natural next question is, *"What characterizes the people with high risk estimates?"* To answer this question, we study the characteristics[15] of the people in the upper tail of the risk distribution of a single shock and compare it to those who are not. We focus on the two main shocks *social_benefits* and *health_expenditures*. Even though this is no causal analysis, the results will give us an insight in the potential over-representation of certain groups in the upper tail of the risk distribution.

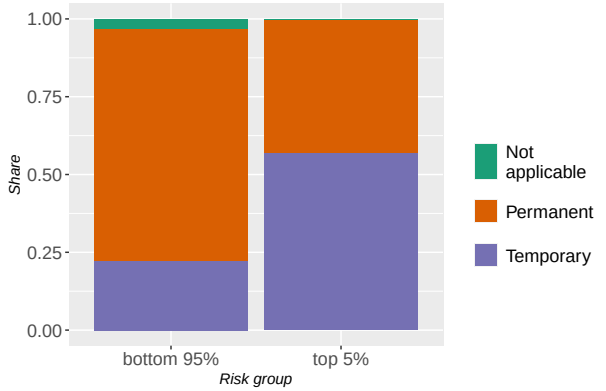
We analyze the distribution of risk estimates for a selection of background variables and the main shocks *social_benefits* and *health_expenditures*. Our primary interest lies in the profile of those individuals in the upper tail of the risk distribution. More specifically, we focus on the features of those individuals with the top 5% highest risk estimates. This percentage roughly corresponds to the order of magnitude of the shock prevalences observed in the sample (see table 2). Figures 5 to 8 show the distribution of characteristics of the individuals in the lowest 95% of the estimated risk distribution versus those in the top 5%. Figures 5 and 6 display the distribution of background features in the domain of employment, education, wealth, and income. Figures 7 and 8 concern more personal characteristics, such as gender, age and marital status.

Contract type, employment type and socio-economic category. Figures 5(a) and 6(a) show the proportion of individuals with a fixed-term contract versus a permanent contract. The share of individuals with a fixed-term contract 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 in the high-risk group (see figs. 5(b) and 6(b)). Lastly, the share of self-employed individuals is substantially smaller in the upper tail of both the labor and health risk distribution, see figs. 5(c) and 6(c). The category 'Other' in these

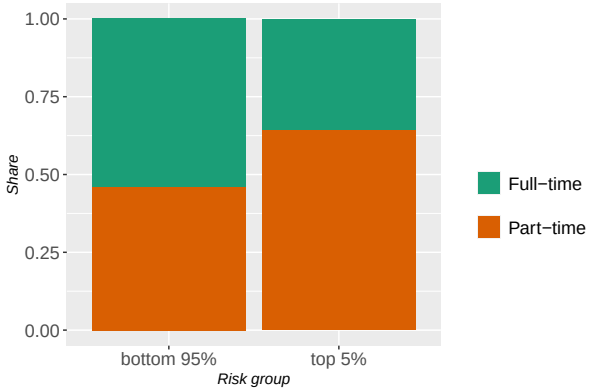
[15] We examine the background characteristics one year prior to the potential occurrence of the shock.

Fig. 5. Risk groups for *social_benefits* shock by employment, income and wealth characteristics

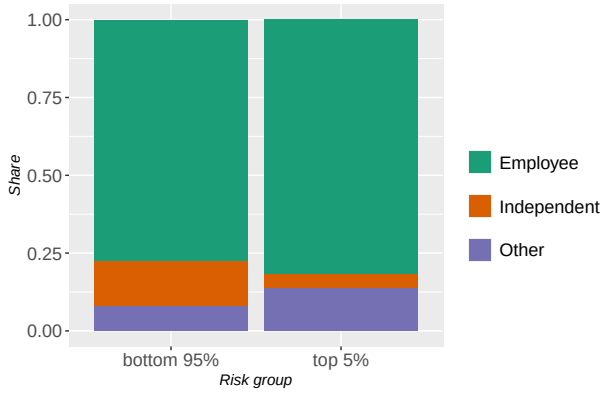
(a) Contract type



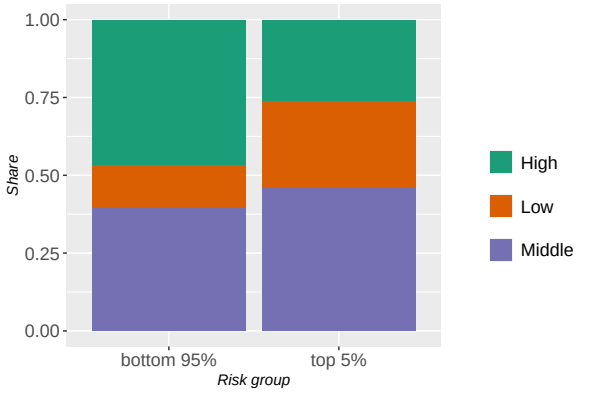
(b) Employment type



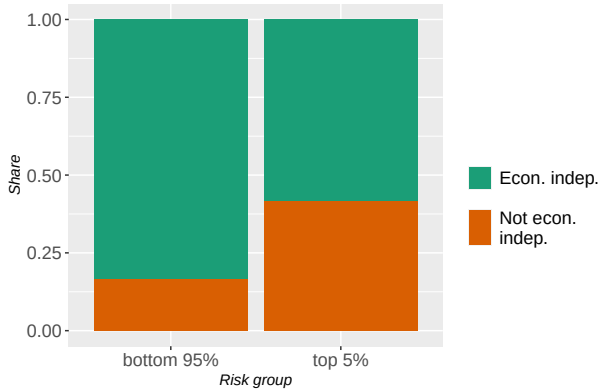
(c) Socio-economic category



(d) Education level



(e) Economic independence



(f) Wealth percentile

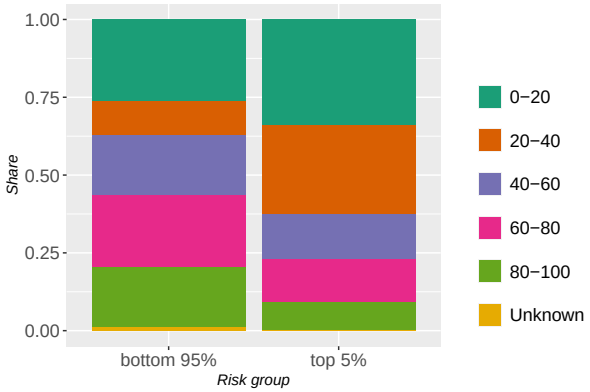
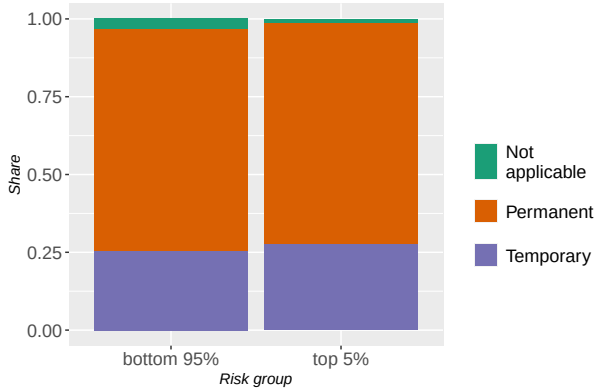
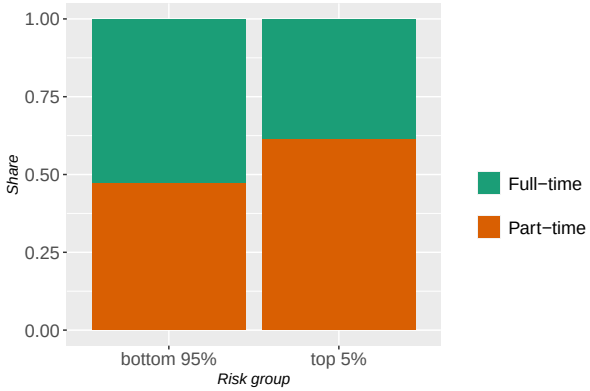


Fig. 6. Risk groups for shock *health_expenditures* by employment, income and wealth characteristics

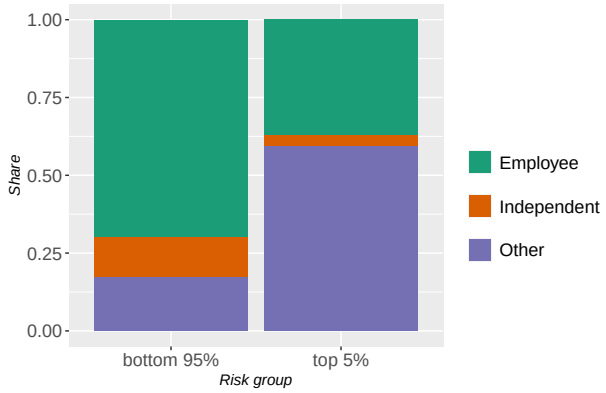
(a) Contract type



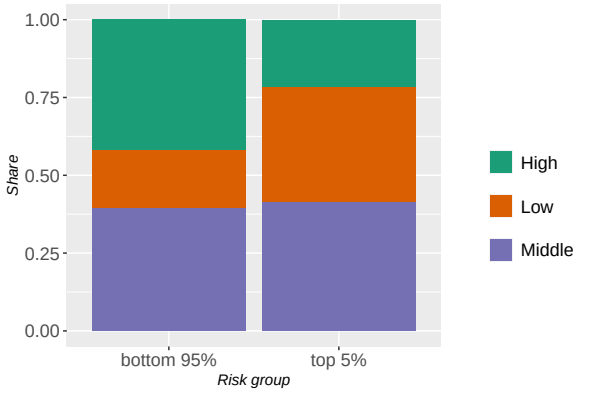
(b) Employment type



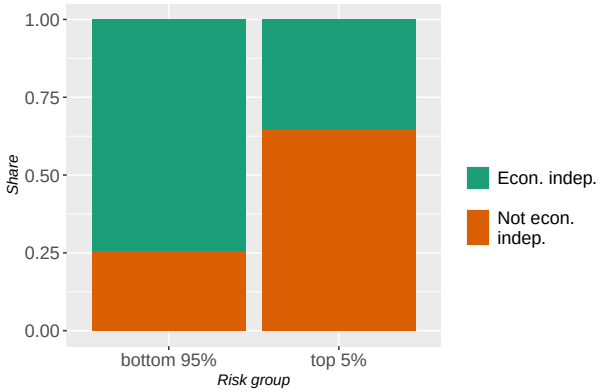
(c) Socio-economic category



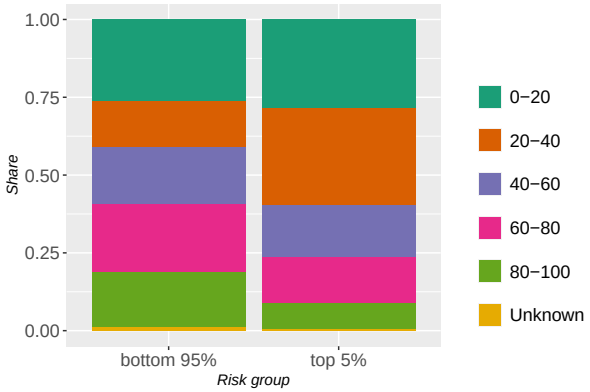
(d) Education level



(e) Economic independence



(f) Wealth percentile



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.

Education level. Figures 5(d) and 6(d) show the distribution of the education level within the low- and high-risk group. For both the labor and health shock, the share of people with a low education level (primary school, lower levels of practical education) is about twice as big in the high-risk tail. The share of individuals with a high education level (university, higher professional education) is about half the size in the upper tail of the labor and health risk distribution compared to the rest.

Economic independence and wealth. Economic independence in this context is defined as having a personal income that is higher than 70% of the minimum (after-tax) 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* compared to the remainder (see fig. 6(e)). Again, because the individuals who are already on social benefits are excluded from the shock *social_benefits*, the level in the bottom 95% is much smaller for *social_benefits* (see fig. 5(e)). Additionally, figs. 5(f) and 6(f) show that there is an over-representation of the 40% individuals with the smallest wealth.[16] The shift is especially substantial for the 20-40th percentile group, and less for the bottom 20% of the wealth distribution. For *social_benefits*, this could again 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. For *health_expenditures*, an explanation could be that younger people are generally less wealthy but also less likely to incur a health shock.

[16]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).

Gender, country of origin and birth cohort. Approximately 60% of the people in the upper tail of the health risk distribution are women (fig. 7(a)). For the labor shock, the imbalance is much smaller (fig. 7(a)). Furthermore, we see an over-representation of individuals born outside the Netherlands in the upper tail of the labor risk distribution (fig. 7(b)). For the health shock, there are slightly fewer people born abroad appearing in the high-risk tail (fig. 8(b)). There is a small over-representation of young people (birth cohort 1983-1992) in the upper risk tail of *social_benefits* (fig. 7(c)). As expected, older people are over-represented in the upper risk tail of *health_expenditures* (fig. 8(c)).

Housing. There are about four times as many individuals with a rental house with rent benefit[17] in the upper tail of the risk distributions of both shocks compared to the rest of the distribution (figs. 7(d) and 8(d)). Albeit to a lesser extent, individuals in rental houses without rent 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 (figs. 7(e) and 8(e)). Looking at the household composition, we see considerably more singles in the upper risk tail of both *social_benefits* and *health_expenditures* (figs. 7(f) and 8(f)). 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*.

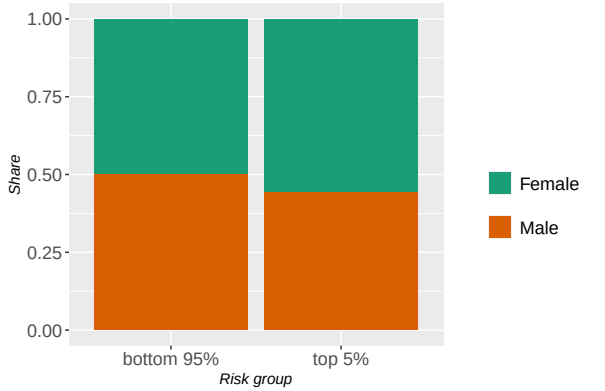
This analysis firmly underscores the unequal distribution of risks within the domains of labor and health across the Dutch population. As illustrated in the preceding figures, specific subgroups significantly dominate the high-risk end of the distribution. Notable among these groups are individuals with temporary employment contracts, those with lower levels of education, individuals originating from outside the Netherlands, residents of rental properties, and single individuals.

A natural extension of this analysis is the exploration of the characteristics shared by individuals who find themselves in the highest risk group for two shocks simultaneously.

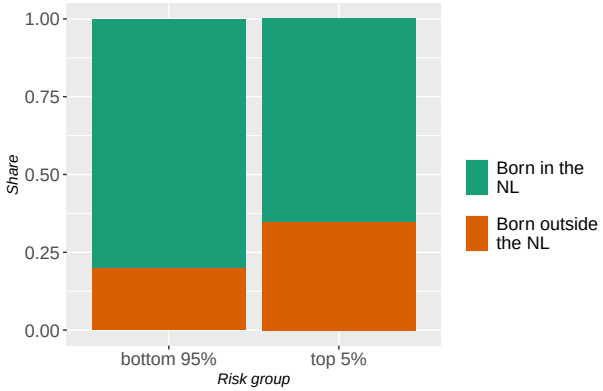
[17] In the Netherlands, rent benefit provides financial assistance to individuals with low incomes who rent their homes.

Fig. 7. Risk groups for shock *social_benefits* by a selection of personal and household characteristics

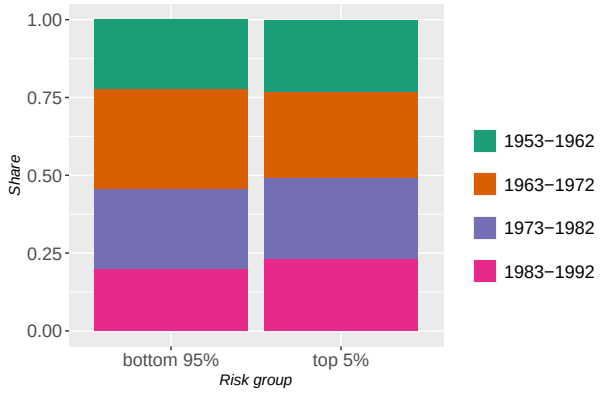
(a) Gender



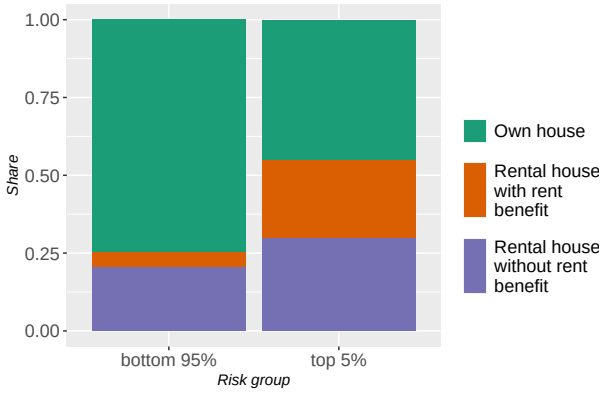
(b) Country of origin



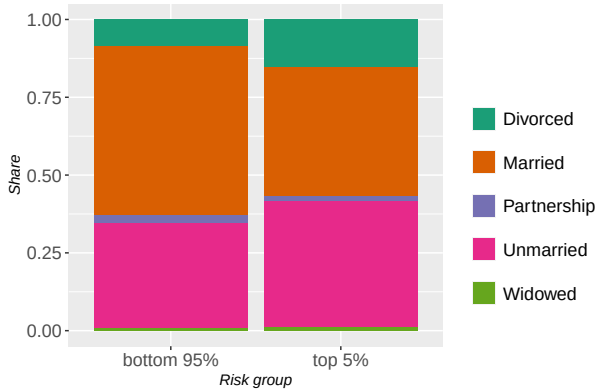
(c) Birth cohort



(d) Housing



(e) Marital status



(f) Household composition

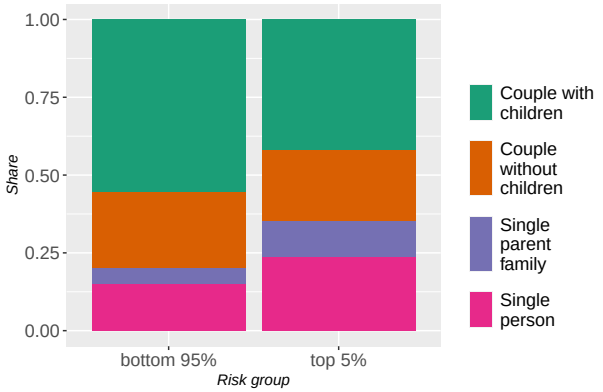
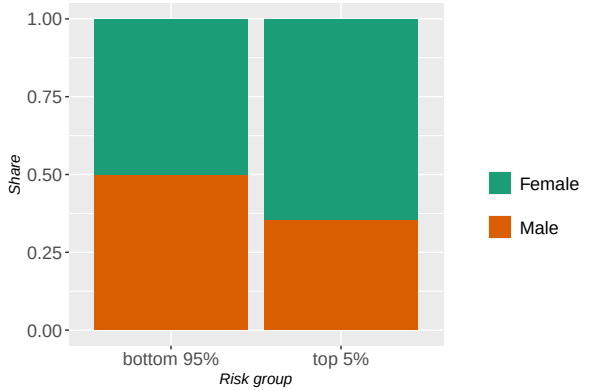
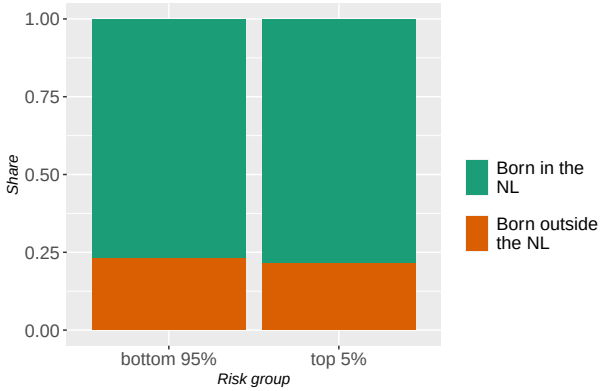


Fig. 8. Risk groups for shock *health_expenditures* by a selection of personal and household characteristics

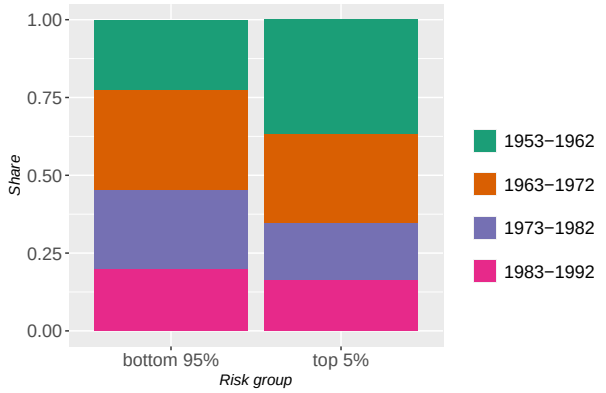
(a) Gender



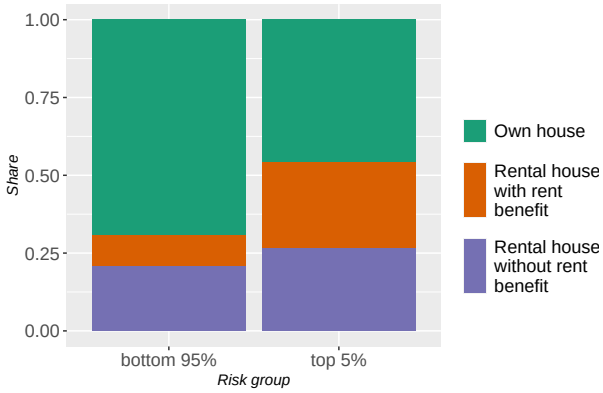
(b) Country of origin



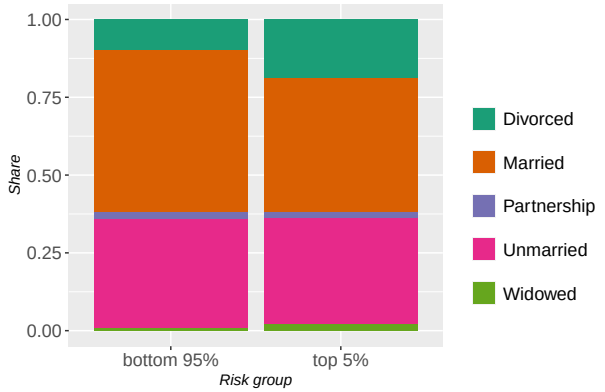
(c) Birth cohort



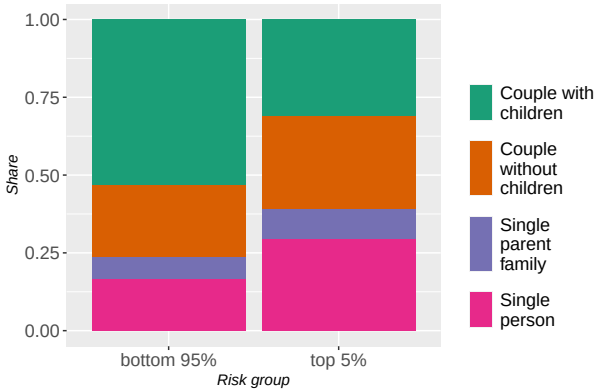
(d) Housing



(e) Marital status



(f) Household composition



The results are displayed in figs. D3 and D4. Notably, the subgroups that stand out in this context exhibit substantial overlap with those identified previously.

One could wonder why we focus on analyzing high-risk individuals rather than those who have already experienced a shock. In principle, our interest lies in understanding the characteristics of people at high risk, not solely those who were unfortunate enough to be affected by a shock. We conducted an analysis of the risk characteristic distribution based on actual occurrences rather than predictions, and the results were largely congruent. In the end, these two groups exhibit large similarities, but this was not evident ex-ante.

Caution should be exercised when interpreting the results in the context of formulating policy implications. The mere over-representation of a particular subgroup in the upper tail of the distribution does not suffice to conclude that this subgroup, on average, faces higher risks than the entire population.[18] Policy makers should be aware that relying solely on these demographic variables for targeting individuals may result in reaching out to a substantial number of individuals who are not genuinely at high risk, leading to a large amount of unnecessary treatments. In section 8 we will look at the potential of targeting specific groups and how conditioning on past shocks might improve the effectiveness of targeting the individuals at high risk significantly.

8. Policy Implications

The results from the previous sections show that it is possible to identify individuals who are at risk of facing setbacks in the domains of labor and health. This knowledge can be used to inform prevention policy where high-risk people receive targeted treatment before the shock occurs. Accurately applied prevention in advance might be a more efficient use of public funds than waiting for predictable adverse events to materialize and only then

[18] For example, in case of a bimodal distribution where there is also a large share of people of that same subgroup with a very low risk estimate.

intervening.[19] For instance, in the context of becoming the recipient of social benefits, prevention policy could entail improving the job market position of those who are vulnerable to losing their job. A similar rationale was applied at the onset of the COVID-19 pandemic, where job retention schemes prevented a surge in unemployment (Scarpetta et al., 2020). However, such broad and untargeted policies are likely to be cost-ineffective. Using ex-ante risk estimates would allow governments to intervene only where needed.

Besides showing that risk selection is possible, the results from the previous sections also show a strong degree of risk concurrence. Individuals who are at risk of facing a particular shock (or who have faced one recently) are also more likely to face another shock in the future. The value of prevention policy is even higher in the presence of risk concurrence, because it not only prevents the specific shock at hand but also potentially decreases the likelihood of subsequent shocks materializing. The merit of social security policy, whether it be of the preventive or restoring type, should then be considered in a wider context to internalize its broader benefits.

The identification of high-risk individuals in this paper either follows directly from a machine learning model that is trained on a rich data set, or it follows from pinpointing which person characteristics (be it socioeconomic indicators or past shock realizations) are strongly correlated with the risk estimates of that machine learning model. Tables 7 and 8 synthesize the degrees to which the three sources of identification can be useful for targeted prevention policy, for the shock *social_benefits* and *health_expenditures* respectively. The first column describes different sub-populations based on risk estimates, socioeconomic characteristics (measured before the shock occurs) that are over-represented in the upper tail of the risk distributions (or combinations thereof) and past shock incidence. The second column shows the prevalence within each sub-population, while the third column shows the percentage of individuals within each sub-population that end up in the upper tail of the risk distribution.[20]

[19] Many other aspects are relevant when considering the feasibility of targeted prevention policies, such as treatment efficacy, timely information production and transparent instrumentation.

[20] Naturally the numbers in the third column are larger than those in the second column, because the prevalence of the shocks is smaller than the five percentiles that make up the upper tail of the risk distribution.

Table 7: Targeting effectiveness of *social_benefits* using various sub-populations.

	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%
single background variables			
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%
multiple background variables			
top 2 background variables	8.7%	31.8%	2.6%
top 3 background variables	14.3%	45.5%	0.5%
all 8 background variables	24.5%	77.5%	0.0%
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%

The fourth column shows the size of each sub-population relative to the total population. To illustrate, individuals with a temporary contract make up 19.3% of the sample, have a 4.5% probability of facing the shock *social_benefits* (compared to 2.0% for the entire population),

Table 8: Targeting effectiveness of *health_expenditures* using various sub-populations.

	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%
single background variables			
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%
multiple background variables			
top 2 background variables	6.4%	5.4%	2.4%
top 3 background variables	6.5%	6.3%	0.7%
all 8 background variables	9.2%	8.1%	0.0%
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%

and 12.9% of them are deemed high-risk.

Ideally one would target a sub-population of manageable size with a high prevalence. It is clear from both tables that the risk estimates of the trained machine learning model are

superior in identifying high-risk individuals. No combination of socioeconomic characteristics and past shock incidence can achieve a higher prevalence for a given targeted population fraction. Single socioeconomic characteristics afford some degree of risk separation, but the targeted population fraction remains prohibitively high. One can combine multiple characteristics to simultaneously increase the prevalence and decrease the size of the targeted sub-population, but the size quickly dwindles as characteristics stack. This highlights that over-representation of a particular socioeconomic characteristic in the upper tail of the risk distribution does not automatically mean that that characteristic can be used to effectively identify high-risk individuals. Past realizations of adverse events on their own achieve similar if not higher levels of prevalence compared to socioeconomic indicators due to the high degree of risk concurrence. Note that not only setbacks from the same domain are informative, but also setbacks from the other domain. Combinations of socioeconomic characteristics and past shock incidence come closest to the first-best achieved by the risk estimates of the trained machine learning model.[21]

While the risk estimates of the trained machine learning model have superior predictive power, they also come at a large cost in terms of data requirements. Section C.2 shows that the risk estimates of a machine learning model trained on a data set with only a few rudimentary variables perform extremely poorly. We only achieve accurate predictions when combining information on the entire Dutch population from many different administrative registries. To implement our predictive model directly as a policy tool would require the continuous collection of large amounts of recent data from many different sources. Data can take time to become available[22], which would render it infeasible for use in next-year predictions, and data is often locked behind strict privacy protection measures.[23] Therefore, policy acting directly on sufficiently accurate individual risk predictions does not seem

[21]The probabilities that our machine learning model outputs can be interpreted as estimates of a latent variable that embodies one's risk exposure. The socioeconomic characteristics and past shock incidence with the highest predictive power are themselves a weaker estimate of that same risk exposure and are correlated with the risk estimate of the machine learning model.

[22]In the case of this paper, the lag is one to two years.

[23]In the case of this paper, the data was only made available to us for the purpose of academic research.

feasible in the near future. Beyond practical considerations, there are doubts whether such policies are desirable and transparent. For example, a trained machine learning model might coincidentally be better at making accurate risk estimates for a particular group of people and worse for other groups. Policy based on those risk estimates would then over-select one group for interventions at the expense of others.[24] In the limit, the most disadvantaged groups of people are likely to not even be present in data sets at all, rendering it impossible to make individual predictions for them.[25]

If the objections to using individual risk estimates weigh stronger than their merit in terms of improved targeting, one might resort to target prevention policies using socioeconomic indicators or past shock incidence. Our results show that this would achieve reasonable levels of risk separation, especially when both are combined. This indicates that a sparse, yet carefully curated, list of variables can capture some of the information that is encapsulated in our rich data set. Van Hoenselaar et al. (2023) and De Klerk et al. (2023) previously noted the predictive value of socioeconomic indicators in the Dutch context. This paper corroborates those findings and shows that past shock incidence is also useful. Both types of information are likely to be available to policymakers already through government registries.

While the predictability of setbacks is not a sufficient condition for targeted prevention policies to be of added value, it is a necessary condition. This paper shows that it is possible to identify high-risk individuals using machine learning and sufficiently rich data. However, on top of this, one would still be in need of cost-effective policy interventions that mitigate risks. While the study of such interventions is beyond the scope of this paper, the results on risk concurrence suggest that interventions targeting multiple vulnerabilities jointly will be most useful.

[24] However, this objection can be raised against any form of targeted prevention policy, regardless of the underlying mechanism that selects who is targeted.

[25] Coverage is near-universal in our data set, but this might not be the case in other applications.

9. Conclusion

In this paper we investigate the risk, distribution and co-incidence of setbacks in the labor and health domains. Our research is made possible by extensive data on millions of Dutch individuals (section 3). Using this data we define multiple shocks in both domains, with each year about one in fifty to one in twenty individuals suffering such shocks (section 4). We then train a gradient boosted tree machine learning model on 2 million individual-year observations and have it predict shock probabilities for another 2 million such observations (section 5). Our work shows that the ex-ante risk of labor and health shocks can be predicted with convincing performance (sections 5.2 and 5.3). The prediction results then allow us to find strong concurrence of risks between setbacks (section 6) and to discover disproportional representations of various socioeconomic groups in the upper risk tails (section 7). However, since the persons at increased risk may only represent a small portion of the identified population groups, group membership alone is not a sufficient predictor for increased risk (section C.2). In section 8 we discuss policy implications of our work. Our main conclusions here are that with enough extensive data it is possible to make dependable risk assessments for labor and health shocks at the individual level, but that actual implementation of targeting based on risk estimates of trained machine learning models is currently infeasible due to practical considerations. A reasonable second-best can be achieved by targeting based on one's membership of socioeconomic groups and one's previous incidence of other setbacks. The latter is a novel insight that derives from our discovery of strong risk concurrence. The presence of domino effects regarding adverse events implies that targeted prevention policies tackling multiple vulnerabilities at the same time are particularly valuable.

This paper leaves various avenues open for future research. Firstly, 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. Secondly, one could investigate the stationarity of the prediction models. Our models are trained using data from a period where the 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. Thirdly, it would be worthwhile to examine which sparse sets of variables achieve reasonable levels of predictive power. This would directly address the concerns regarding data requirements that are mentioned in section 8. The authors of this paper intend to continue by studying whether the recovery from setbacks is also predictable ex-ante. Combining that with the insights of this paper would give a more complete picture of individual resilience. It would enrich the policy insights that we can provide, since policy should ideally be targeted 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.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

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.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5).

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.

Scarpetta, S., Pearson, M., Hijzen, A., and Salvatori, A. (2020). Job retention schemes during the COVID-19 lockdown and beyond. *OECD Policy Responses to Coronavirus (COVID-19)*.

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

Appendix A. Details Shock Definitions

This overview gives more background information 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 or 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 euro. 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.

Appendix B. *LightGBM* package 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.

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.

Besides the package parameters there is one other aspect of performance worth mention-

ing, 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.

Appendix C. Supplementary Analysis of Predictions

C.1. Variable Importance

Although it is difficult to draw any conclusions from the trained model itself, as previously mentioned, a cursory glance can provide some ostensible insights. The trained model reports a *variable importance* that attempts to express the relative significance of a variable for predicting outcomes. The main caveat of the variable importance is that it often misattributes importance to categorical variables or when variables are correlated to other variables. Therefore, it is prone to portraying a skewed ranking.

When we look at the top 25 variables with the highest importance for the 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 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 as 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.

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 could be an indicator of what we show later on; that shocks in the two domains are indeed significantly correlated.

C.2. Predictions with Limited Variables

A highly policy-relevant question is whether we need all of our many variables to make accurate predictions. A desirable alternative would be to use only easily accessible background characteristics. We tested this by making predictions for the *social_benefits* shock

Table C1: 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}	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

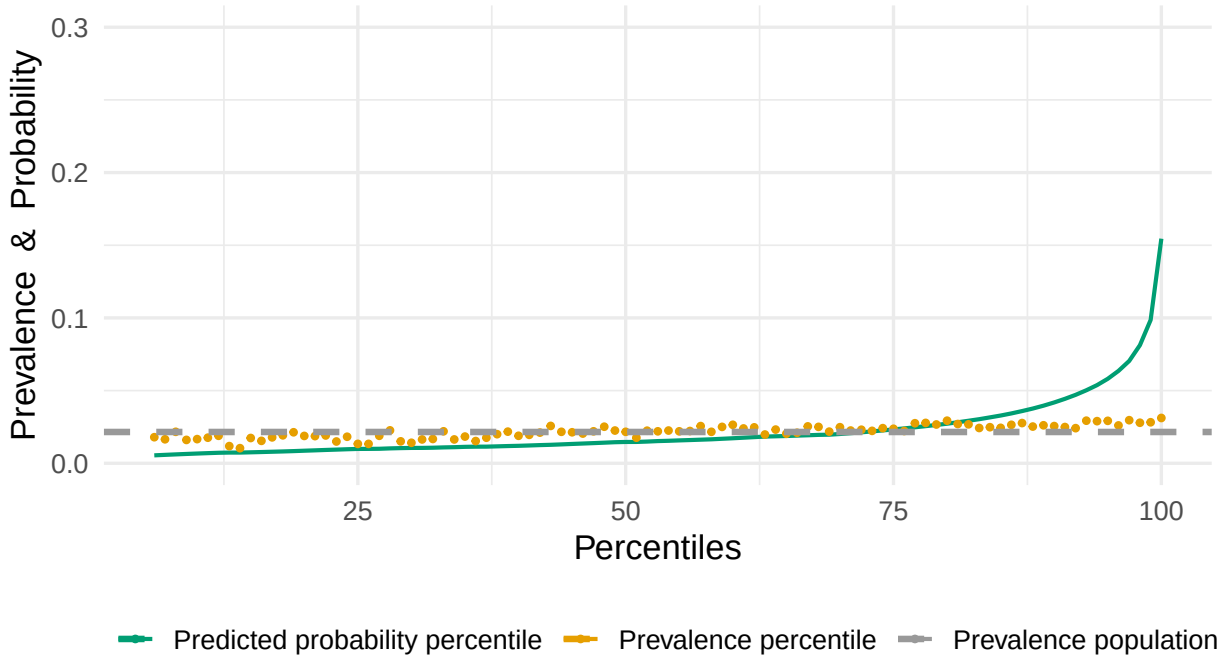
with a model trained on only 11 characteristics, shown in table C2. These 11 characteristics comprise the most rudimentary data that the Dutch government has access to, through its Personal Records Database (*Basisregistratie Personen* in Dutch).

Table C2: Background characteristics that are easily accessible for governments.

	Birth year	Country of birth	Mother's country of birth
Gender	Birth month	Marital status	Father's country of birth
ZIP code	Municipality	Housing status	Household composition

Figure C1 shows the prevalence of shock realizations per risk estimate bin, following the method from section 5.3. Unlike in figs. 1 and 2, the prevalences do not trace the risk estimates at all and instead roughly follow the population average. This means that model performance is very poor, which is further emphasized by the poor AUC of 0.55 (0.5 represents random guessing) and poor F1-score of 0.05.

Fig. C1. Realization prevalence and risk estimates for shock *social_benefits* when trained on only 11 prominent background characteristics.



The conclusion that we can draw, however, is significant. The high degree of performance found in sections 5.2 and 5.3 cannot be reproduced with a severely limited variable set. Apparently, these background characteristics are not sufficient proxies for the plethora of other variables in our data set. While our other results show that groups with high shock probabilities can be successfully identified, merely selecting on broad categories such as ZIP code or country of birth is insufficient to reproduce such results.

C.3. Predictions with Oversampling

One pervasive issue that we have mentioned is the low prevalence rate of shocks. As we saw in section 5.2, the low prevalence rate distorts the significance of for example the accuracy statistic. Additionally, there is the possibility that our models do not optimally predict shocks because relatively few realizations were seen in the training set.

One way to mitigate this issue is by oversampling shock realizations. By curating the

training data we can artificially increase the prevalence of shocks. Because we have such 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 shock *social_benefits* to have twice the normal prevalence.

Fig. C2. Realization prevalence and risk estimates for shock *social_benefits* when trained on oversampled data with twice the prevalence. Also shows reference prevalence.

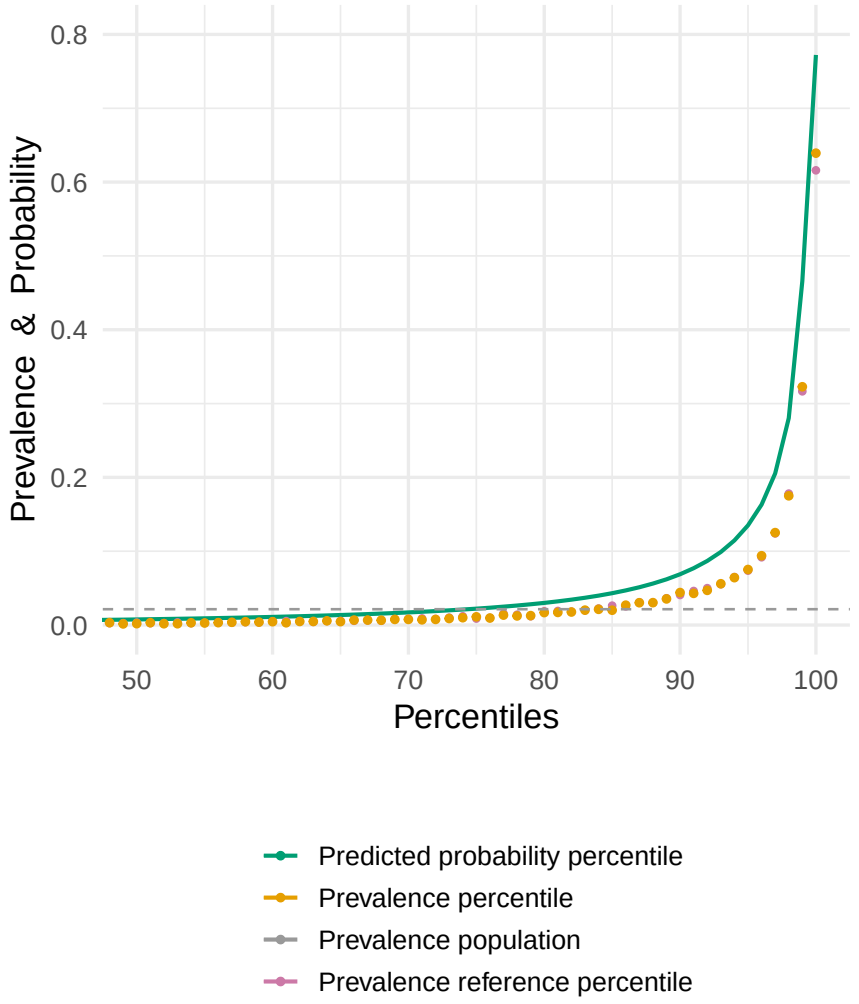


Figure C2 shows the prevalence and risk estimates for each percentile bin of predicted probabilities. What stands out is that the risk estimates are consistently higher than the observed prevalences for each bin, which can be attributed to the fact that this model was trained to expect twice the prevalence that it encounters in the test set.

More importantly, we find that this model is better at predicting in the high probability tail. Figure C2 shows a reference prevalence per bin resulting from a model trained on data without oversampling. In percentiles with the highest risk estimates according to the model trained on oversampled data, the prevalence of shock realizations is higher than the reference prevalence. Apparently, because this model has seen more shock realizations during training, it has become better at successfully discerning groups with high shock probabilities. The trade-off seems to be that this model is less proficient at accurately predicting in the bottom half of percentiles. This suggests that during training the model can be geared towards a specific application, such as identification of the upper risk tail.

C.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 below. 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 6 to investigate how different risks relate to each other. Focusing on joint shock definitions would rule out such analysis.

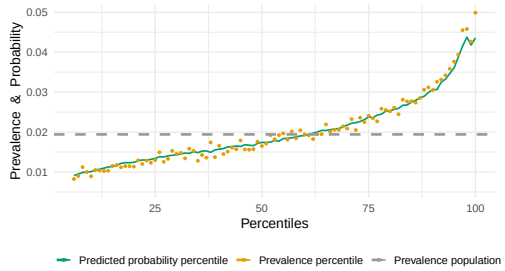
We have also experimented with the time horizon of the shock definitions. Instead of requiring that shocks occur in a given year, we allowed them to occur at any point in a period of two or more years. This could improve predictive power because the trained prediction model has to be less precise about 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.

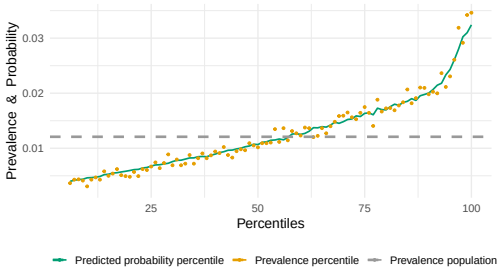
Appendix D. Additional Figures and Tables

Fig. D1. Joint risk distribution of *social_benefits* (on x-axis) and various health shocks (on y-axis).

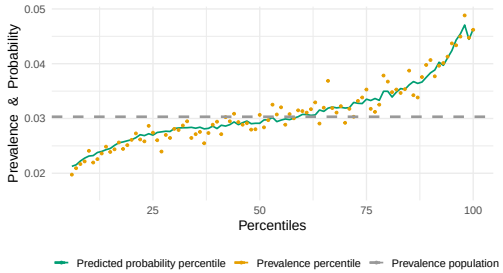
(a) *mental_health_medication*



(b) *mental_health_treatment*



(c) *physical_health_treatment*



(d) *physical_health_ic*

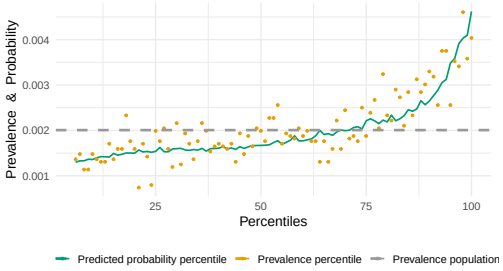


Fig. D2. Joint risk distribution of *health_expenditures* (on x-axis) and various labor shocks (on y-axis).

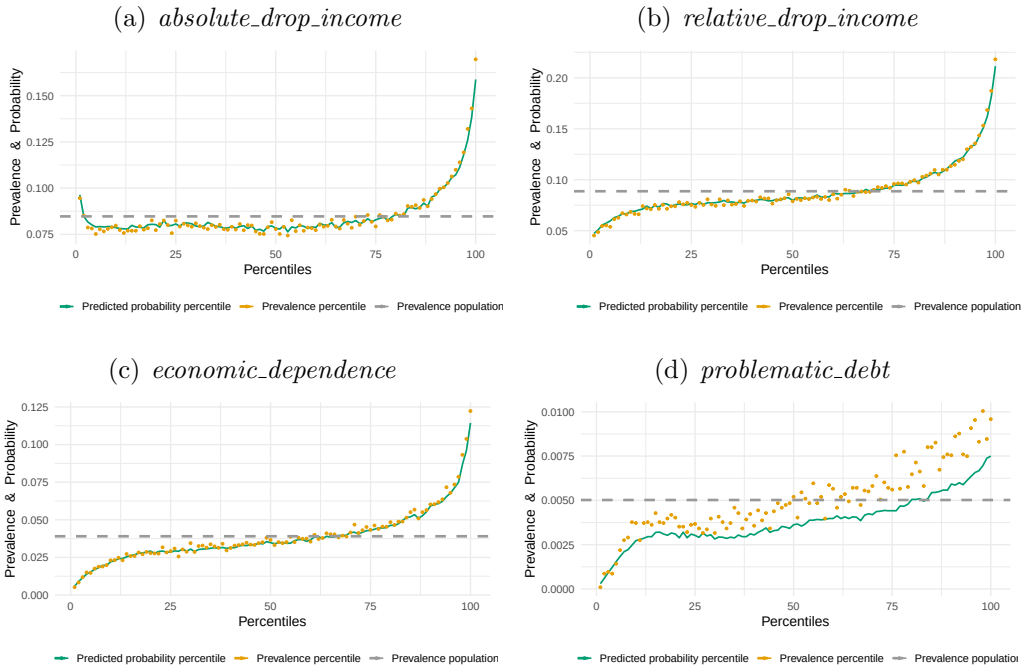
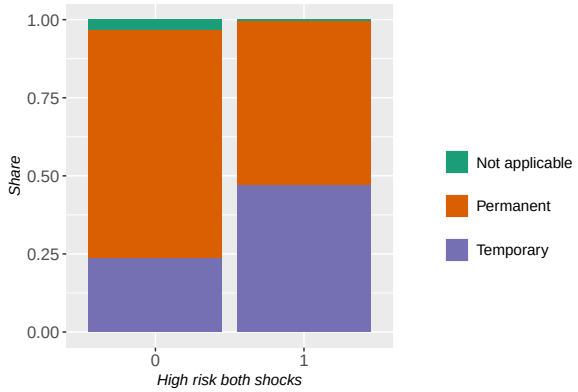


Table D1: This table displays the multiplication factor of the prevalence of a shock in year t for individuals that experienced a different shock in year $t-1$, relative to the unconditional prevalence of experiencing that shock in year t.

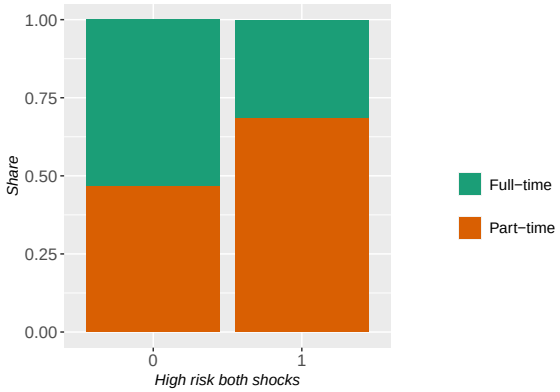
		conditional on shock in t-1:											
shock in t:	unconditional prevalence (%)	*social_benefits*	*relative_drop_income*	*absolute_drop_income*	*problematic_debt*	*economic_dependence*	*health_expenditures*	*physical_health_expenditures*	*physical_health_treatment*	*physical_health_ic*	*mental_health_expenditures*	*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

Fig. D3. Risk groups for shock *social_benefits* and *health_expenditures* 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.

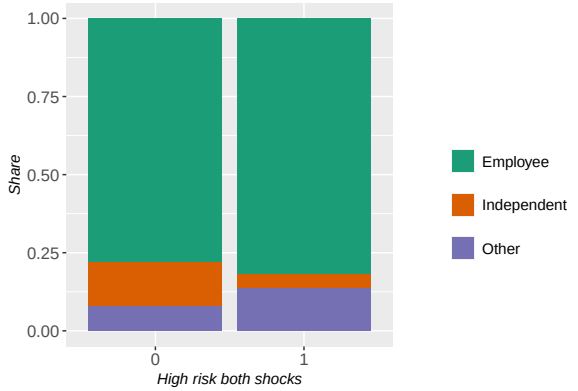
(a) Contract type



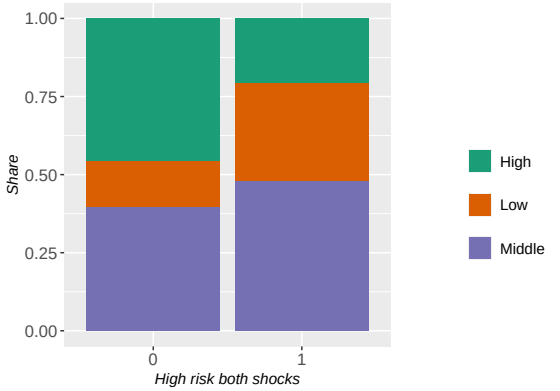
(b) Employment type



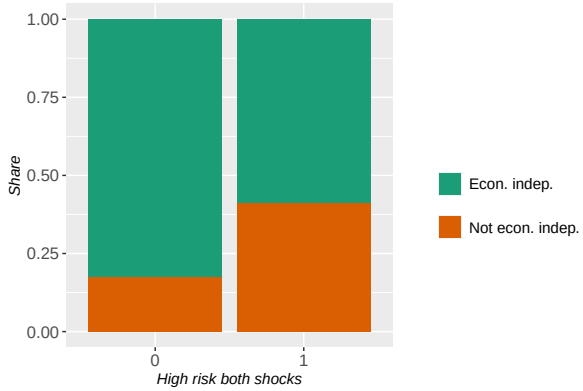
(c) Socio-economic category



(d) Education level



(e) Economic independence



(f) Wealth percentile

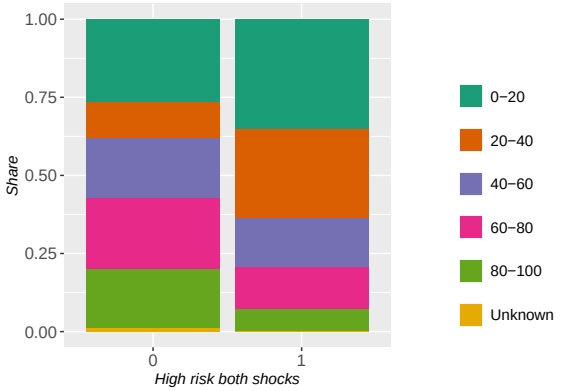
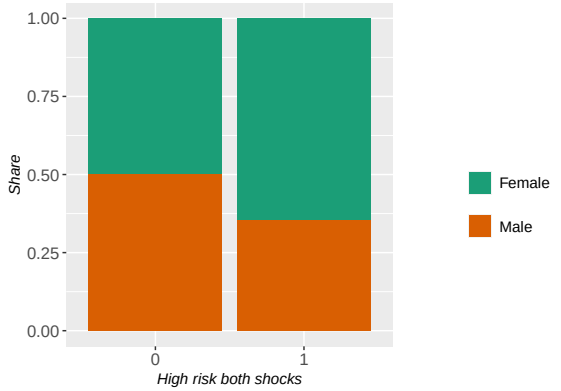
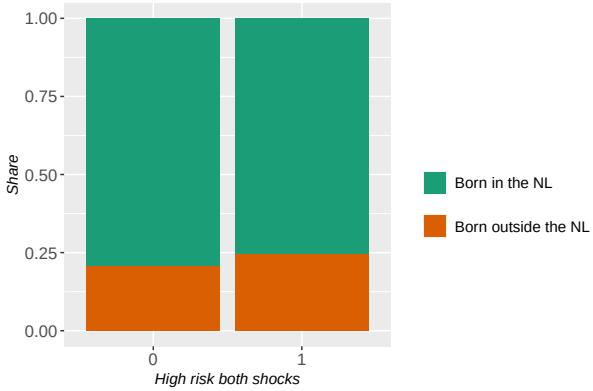


Fig. D4. Risk groups for shock *social_benefits* and *health_expenditures* by a selection of personal and household characteristicsIndividuals who are in the top 5% of the risk distribution of both shocks are considered at high risk for both shocks.

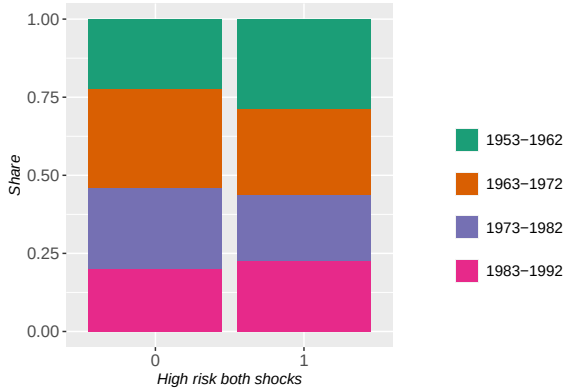
(a) Gender



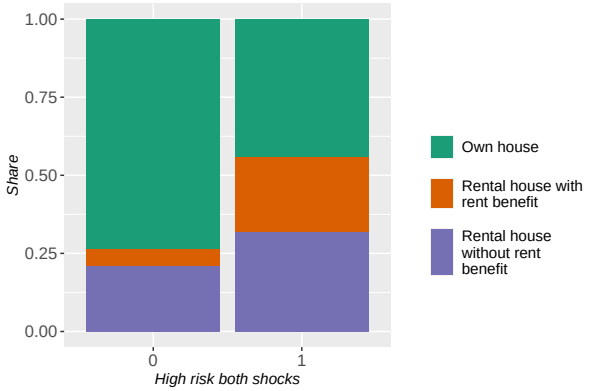
(b) Country of origin



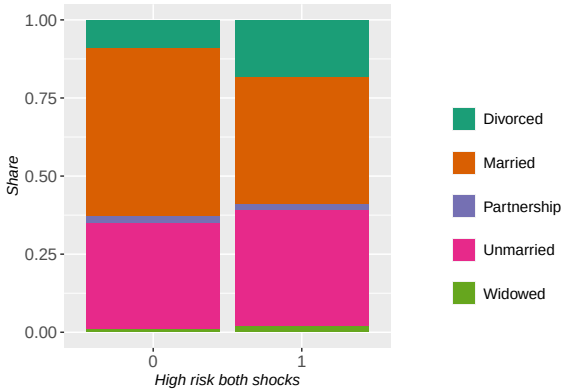
(c) Birth cohort



(d) Housing



(e) Marital status



(f) Household composition